DOCUMENT RESUME

ED 275 162 FL 016 066

AUTHOR Culhane, Terry, Ed.; And Others
TITLE Practice and Problems in Language Testing. Papers from the International Symposium on Language Testing (7th, Colchester, England, 1984). Occasional Papers, No. 29.
INSTITUTION Essex Univ., Colchester (England). Dept. of Language and Linguistics.
PUB DATE 84
NOTE 198p.; For individual papers, see ED 227 652 and FL 016 067-079.
AVAILABLE FROM Department of Language and Linguistics, University of Essex, Wivenhoe Park, Colchester, C04 3SQ, England (3 British pounds).
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS Children; Classroom Communication; Cloze Procedure; Communicative Competence (Languages); Elementary Secondary Education; *English (Second Language); English for Science and Technology; English for Special Purposes; Grammar; Higher Education; Immigrants; *Language Tests; Listening Comprehension; Reading Comprehension; Self Evaluation (Individuals); *Test Construction; *Testing Problems; Test Validity; Vocabulary Development; Writing Evaluation

ABSTRACT

Papers include: "ESP Tests: The Problem of Student Background Discipline (J. C. Alderson, A. H. Urquhart); "The Lexical Richness of L2 Written Productions and the Validity of Vocabulary Tests" (P. J. L. Arnaud); "The Construction of a Modular Thematic Task Bound English Language Achievement Test Battery for Kuwait" (M. I. Baghdady); "The Fill-In Test: A Modified Multiple-Choice Cloze Technique to Test Reading Comprehension of English as a Foreign Language (M. Bensoussan, R. Ramraz); "Self-Assessment in English as a Foreign Language: Some Preliminary Ideas" (N. A. Berkoff); "Testing Listening Comprehension: Notions and Functions as 'Discrete Points' in a Communicative Context" (M. Boeijen); "Can Grammatical Correctness and Communication Be Tested Simultaneously?" (B. Falk); "Advance Prediction of Difficulty with C-Tests" (C. Klein-Braley); "A Regression Analysis of Direct and Indirect Measures of English as a Second Language Writing Compositions" (K. Perkins); "The Factorial Validity of C-Tests" (U. Raatz); "Language Tests as Indicators of Academic Achievement" (P. M. Rea); "Input and Output in Language Testing" (E. Shohamy); "Language Testing with Immigrant Children" (E. Sussmilch); and "'Classroom Language'--A Neglected Area in Foreign Language Teaching and Testing" (B. Voss). (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED275162

University of Essex

Department of
Language and Linguistics

Occasional Papers

Practice and problems in language testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Culhane

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IUS

edited by
Terry Culhane
Christine Klein Braley
Douglas K. Stevenson

No.29

FL 016 066

Occasional Papers

29: Practice and Problems in Language Testing

edited by Terry Culhane,
Christine Klein Braley and Douglas K. Stevenson

Published by the Department of Language and Linguistics
Printed by the University of Essex, Wivenhoe Park, Colchester, CO4 3SQ
Autumn 1984
Price: £3.00 post free

PRACTICE AND PROBLEMS IN LANGUAGE TESTING

CONTENTS

CONTENTS (i)

FOREWORD (ii)

ESP TESTS: THE PROBLEM OF STUDENT BACKGROUND DISCIPLINE 1
J C Alderson (Lancaster) and A H Urquhart (Aston)

THE LEXICAL RICHNESS OF L2 WRITTEN PRODUCTIONS AND THE VALIDITY OF VOCABULARY TESTS 14
P J L Arnaud (Lyon 2)

THE CONSTRUCTION OF A MODULAR THEMATIC TASK BOUND ENGLISH LANGUAGE ACHIEVEMENT TEST BATTERY FOR KUWAIT 29
M I Baghdady (Exeter)

THE FILL-IN TEST: A MODIFIER MULTIPLE-CHOICE CLOZE TECHNIQUE TO TEST READING COMPREHENSION OF ENGLISH AS A FOREIGN LANGUAGE 44
M Bensoussan (Haifa) and R Ramraz (Haifa)

SELF-ASSESSMENT IN ENGLISH AS A FOREIGN LANGUAGE: SOME PRELIMINARY IDEAS 66
N A Berkoff (Hebrew, Jerusalem)

TESTING LISTENING COMPREHENSION: NOTIONS AND FUNCTIONS AS "DISCRETE POINTS" IN A COMMUNICATIVE CONTEXT 76
M Boeijen (Cito, Arnhem)

CAN GRAMMATICAL CORRECTNESS AND COMMUNICATION BE TESTED SIMULTANEOUSLY? 90
B Falk (Utrecht, The Netherlands)

ADVANCE PREDICTION OF DIFFICULTY WITH C-TESTS 97
C Klein-Braley (Duisburg)

A REGRESSION ANALYSIS OF DIRECT AND INDIRECT MEASURES OF ENGLISH AS A SECOND LANGUAGE WRITING COMPOSITIONS 113
K Perkins (Southern Illinois, USA)

THE FACTORIAL VALIDITY OF C-TESTS 124
U Raatz (Colchester)

LANGUAGE TESTS AS INDICATORS OF ACADEMIC ACHIEVEMENT 140
P M Rea (Dar es Salaam)

INPUT AND OUTPUT IN LANGUAGE TESTING 159
E Shohamy (Tel Aviv)

LANGUAGE TESTING WITH IMMIGRANT CHILDREN 167
E Süßmilch (Düsseldorf)

'CLASSROOM LANGUAGE' - A NEGLECTED AREA IN FOREIGN LANGUAGE TEACHING AND TESTING 177
B Voss (Bielefeld)

Foreword

For its seventh international symposium on language testing the IUS once again had the privilege and pleasure of making use of the facilities of the University of Essex in Colchester. The papers presented there, which are reprinted in this volume, reflect the practical and theoretical interests of scholars from four different continents and a variety of educational and government institutions.

Language testers are sometimes accused of being so fascinated with the 'number-crunching' aspects of their discipline that they forget to consider the human and societal aspects of their research. However it seems to us that the contributors collected here document that this accusation is unfounded. The authors' overriding concern is that of fairness in measurement procedures, and this, we feel, reflects what our discipline is really all about.

Our thanks are due to the Vice-Chancellor of Essex University and his staff who made us so welcome; to the British Council who once more sponsored the symposium through their English Language Testing Service. Finally, to Sue Tucker and Noreen Proudman who handled much of the correspondence.

The Editors

ESP TESTS: THE PROBLEM OF STUDENT BACKGROUND DISCIPLINE

J C Alderson (University of Lancaster) and A H Urquhart (University of Aston)

This paper reports on three studies designed to investigate the effect of students' background discipline - i.e. knowledge of a particular academic area - on their comprehension of written texts. One of the principal motivations underlying the studies was the wish to collect empirical data on the necessity or desirability of ESP proficiency tests for tertiary level students coming to study in an English-speaking country.

The studies were as follows:

a) Study 1: the pilot study; at Aston University; five short texts gap-filling task.

b) Study 2: at Aston and Lancaster Universities: same five texts; gap-filling and questions/short answers.

c) Study 3: at Aston and Lancaster Universities; different versions of the British Council ELTS Study Skills Module (MI).

Study 1: An account of this study has already appeared in print (Alderson and Urquhart 1983). The description given here will therefore concentrate on those parts of the study which also relate to Study 2.

Subjects: The students taking part were studying on an English for Study Skills Course at the University of Aston. They fell into four groups depending on the academic area in which they were going to pursue their post-graduate studies. The groups were 1) Development Administration/Finance (15 students); 2) Engineering (11); 3) Maths and/or Physics (6); 4) Liberal Arts (5).

According to an internal placement test, the first three groups were of very similar linguistic proficiency; the fourth group was more proficient.

Texts: There were five texts in all, two Engineering, two Development Admin/Finance, and one 'general' text. The Engineering texts (titles have been supplied for reference purposes) were 'Turbines', from an article in an Engineering periodical (Hulme, 1981) and 'Electrolytes', from an academic monograph (Gregory 1972). The Development/Finance texts were 'Malaysia' and 'Polanyi', both from a university text-book (Latham 1978). The 'general' text was 'Quixote', one of the 'Rate-builder' texts from the SRA 3B kit (Science Research Associates 1963). According to the publishers, this kit is designed for high or junior-high school L1 students.

Each text was between 250 and 280 words long. 'Electrolytes' included a diagram with caption, and 'Malaysia' a table. The Fog index for each text was between 17.5 and 18.5.

Tasks: A number of lexical items (between 15 and 22) were deleted from each text. In the judgement of the authors, all were restorable in terms of the text itself, without, that is, appeal to knowledge outside the text. An effort was made to choose items whose restoration depended on an overall grasp of the text, rather than on nearby cotextual items. In the case of 'Electrolytes' and 'Malaysia', certain items could be restored only after consulting the diagram or table respectively.

Given that the object of the experiment was to test the effect of background knowledge, it might appear paradoxical to insist that all items should be restorable on the basis of the text. Anything else, however, would have been trivial. The ability to restore the missing item in the text, "The battle of Hastings took place in the year ..." clearly depends on background knowledge, but in an uninteresting way. It would seem that one of the assumptions underlying traditional 'general' tests is that the information required to be extracted from the text is equally available to all readers. Since it is one of the major hypotheses of this experiment that in spite of such availability, certain groups will still be advantaged by a particular text, it is clearly essential that in the experiment, all the required information should at least in theory be equally accessible to all subjects.

Results: Detailed results can be found in Alderson and Urquhart (op.cit). They can be summarized here as follows:

1) on the two Engineering texts, the Engineering group did better than the Development Admin/Finance group.

2) similarly, the Development Admin/Finance group outperformed the Engineers on the two Development texts.

3) there was a marked similarity between the performance of the Engineers and Physics/Maths students, on the one hand, and the Development and Liberal Arts groups on the other. The latter two groups outperformed the Engineering and Physics/Maths group on 'Quixote', the 'general' text, suggesting that Engineering and Science students may be disadvantaged by such texts.

4) there was a marked text effect. Overall, 'Turbines' proved to be the easiest text, and 'Polanyi' and Quixote' the most difficult. There was no reason to predict this: the only measure of linguistic complexity used, the Fog Index, had indicated that all five texts were all at the same level. At this stage, one should perhaps refer, not to text difficulty, but to text x method. Indeed, without a satisfactory correlative of linguistic complexity, this is perhaps all that can ever be claimed. The text effect did not affect the hypothesis; it may, however, be relevant to the interpretation of the results of Study 2, as discussed below.

Study 2: This fell into two parts, namely:

1) an exact replication of Study 1
2) a partial replication using the same texts but a different test task, namely Short Answers.

The reason for replicating the Pilot Study is obvious - the numbers involved there were very small, too small, in fact, to justify calculating significance values. The reason for extending the Pilot Study by adding another test task is also obvious: without such an extension, the study is vulnerable to method effect.

Subjects: students on ESS courses at both Aston and Lancaster University took part. The groups were very similar in academic discipline to those participating in Study 1, namely:

1) Development Admin/Finance (Aston) and Economics (Lancaster); from now on referred to as DAFE.

2) Engineers (ENG)

3) Science and Maths (S&M)

4) Liberal Arts (L/A)

Numbers of subjects doing (a) gap-filling (b) Short Question/Answer, and (c) both are shown in table 1.

Table 1

	Gap-filling			Short Answer			Both	
	Aston	Lancaster	T	Aston	Lancaster	T	Both	
DAFE	18	8	26	24	9	33	16/8	24
ENG.	17	4	21	16	4	20	16/4	20
S M	32	6	38	28	6	34	24/6	30
L/A	13	9	22	10	16	26	10/9	19

The Gap-filling test was administered first; the Short Answer test about a week later.

Results: The Gap-filling test was marked three times, once for exact restoration, once for exact restoration, once for synonyms, and once for any response judged acceptable in the general context. Results are presented in Table 2.

TABLE 2

A. Gap-filling (Variable N)

		Economics						Engineering								
		Polanyi			Malaysia			Turbines			Electrolytes			Quixote		
		a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
DAFE	X	9	31	43	17	33	39	33	47	56	19	34	40	10	28	33
	SD	9	23	22	8	16	22	14	20	21	12	18	16	9	18	20
ENG	X	3	14	27	10	23	24	32	49	52	21	40	47	5	18	25
	SD	4	12	18	9	17	19	17	19	19	12	16	16	7	12	14
S&M	X	6	17	27	20	31	33	30	42	47	28	47	54	7	20	22
	SD	6	13	17	7	11	14	15	21	20	11	13	15	6	11	14
L/A	X	13	30	54	21	33	34	43	61	63	22	37	44	16	38	44
	SD	12	19	24	14	23	21	13	17	23	16	21	21	11	20	21

a= exact

b= synonym

c= acceptable

Table 3
Gap-filling: significant differences between means (t-test) a = exact b = synonym c = acceptable

1) Polanyi	ENG			S&M			L/A		
	a	b	c	a	b	c	a	b	c
DAFE	NS	$p<.05$	$p<.05$	NS	$p<.05$	$p<.01$	NS	NS	NS
ENG				NS	NS	NS	$p<.05$	$p<.05$	$p<.01$
S&M							$p<.05$	$p<.05$	$p<.001$

2) Malaysia	ENG			S&M			L/A		
	a	b	c	a	b	c	a	b	c
DAFE	NS	NS	$p<.05$	NS	NS	NS	NS	NS	NS
ENG				$p<.01$	NS	NS	$p<.05$	NS	NS
S&M							NS	NS	NS

3) Turbines	ENG			S&M			L/A		
	a	b	c	a	b	c	a	b	c
DAFE	NS	NS	NS	NS	NS	NS	NS	NS	NS
ENG				NS	NS	NS	NS	NS	NS
S&M							$p<.05$	$p<.05$	$p<.05$

4) Electrolytes	ENG			S&M			L/A		
	a	b	c	a	b	c	a	b	c
DAFE	NS	NS	NS	NS	$p<.05$	$p<.01$	NS	NS	NS
ENG				NS	NS	NS	NS	NS	NS
S&M							NS	NS	NS

5) Quixote	ENG			S&M			L/A		
	a	b	c	a	b	c	a	b	c
DAFE	NS	NS	NS	NS	NS	$p<.05$	NS	NS	NS
ENG				NS	NS	NS	$p<.01$	$p<.01$	$p<.01$
S&M							$p<.01$	$p<.01$	$p<.001$

Significant differences between groups for all three methods of scoring are presented in Table 3.

The following points emerge:

1) The DAFE group did significantly better on the Development texts than did the Engineers, although on 'Malaysia', there was no difference between DAFE and S&M.

2) However, on the Engineering texts, the Engineers did not outperform the DAFE group, although S&M did significantly better on 'Electrolytes'.

3) L/A did significantly better than either S&M or ENG or 'Quixote'.

Results for the Short Answer test are presented in Table 4.

Table 4

B. Short Answers

		Polanyi	Malaysia	Turbines	Electrolytes	Quixote
DAFE	X	47.1	56.1	53.8	42.0	55.4
	SD	25.1	22.3	17.1	28.7	22.0
ENG	X	24.5	42.3	58.1	45.1	48.5
	SD	22.7	23.7	24.1	25.9	21.9
S&M	X	34.9	46.1	58.0	47.5	40.8
	SD	22.3	20.P	16.4	18.6	21.3
L/A	X	55.7	54.4	60.3	46.6	70.8
	SD	22.8	20.0	20.2	19.4	14.3

Significant differences between groups are shown in Table 5. (See over).

The pattern of results is similar to that shown by the scores on Gap-filling. On the two Development texts, DAFE performed significantly better than ENG. However, there was no significant difference between DAFE and ENG on the two Engineering texts. The L/A group outperformed DAFE and ENG on 'Polanyi' and all other groups on 'Quixote'.

If one compares results under the two test methods, the following points emerge:

1) there was a strong method effect. Although the overall correlation between scores on both tests is quite high (.78), correlations between scores on particular tests vary widely. Another sign of method effect is that the rank order of text difficulty is not the same under both methods. On Gap-filling, 'Electrolytes' is a comparatively easy text, ranked second overall; 'Quixote' is difficult, ranked fourth. On Short Answers, however, these ranks are reversed.

Table 5

Significant differences between means (t-test)

B. Short answers

1) Polanyi	ENG	S&M	L/A
DAFE	p .01	NS	NS
ENG		NS	p .001
S&M			p .01
2) Malaysia	**ENG**	**S&M**	**L/A**
DAFE	.05	NS	NS
ENG		NS	NS
S&M			NS
3) Turbines	**ENG**	**S&M**	**L/A**
DAFE	NS	NS	NS
ENG		NS	NS
S&M			NS
4) Electrolytes	**ENG**	**S&M**	**L/A**
DAFE	NS	NS	NS
ENG		NS	NS
S&M			NS
5) Quixote	**ENG**	**S&M**	**L/A**
DAFE	NS	p .05	p .01
ENG		NS	p .001
S&M			p .001

2) In spite of this, the text effect found in Study 1 remains strong over both methods, with 'Turbines' being the easiest text, and 'Polanyi' the most difficult.

3) As in Study 1, the ENG and S&M groups performed very similarly (though not identically). This was also true of DAFE and L/A. This finding, of course, may have an important bearing on the number of ESP tests required.

4) The main point that emerges under both methods is that the hypothesis was only partially confirmed: the DAFE group outperformed the Engineers on Development texts, and the L/A group did better on 'Quixote' than any other group. However, the Engineers did not appear to be advantaged by reading Engineering texts.

In part, this may be due to differing degrees of linguistic proficiency. Pre-testing of students at both universities revealed that the L/A group was the most proficient. At Aston, the three other groups were at the same level, but the Lancaster Economists were superior both to ENG and S&M.

Linguistic proficiency may go part-way to explaining the results. For the Lancaster students, correlations were available between scores on different texts, and scores on ELBA and ELTS tests. These correlations are set out in Table 6.

Table 6

Test correlations with ELBA and ELTS

A. Short Answer	N=35			N=20		
	ELBA 1	ELBA 2	ELBA Total	ELTS G1	ELTS G2	ELTS M1
Malaysia	.45	.38	.43	.45	NS	.38
Polanyi	.52	.52	.56	.68	.69	.69
Turbines	.42	.50	.50	.55	.41	.49
Electrolytes	NS	NS	NS	.44	NS	NS
Don Quixote	NS	.45	.40	.61	NS	.49

B. Gap-filling (Acceptable)	ELBA N 21/25			ELTS N = 11/13		
	1	2	Total	G1	G2	M1
Malaysia	NS	.34	NS	.55	NS	.54
Polanyi	.53	.61	.63	.67	.51	.71
Turbines	.57	.66	.66	.73	NS	NS
Electrolytes	NS	NS	NS	NS	NS	NS
Don Quixote	.50	.46	.51	NS	NS	NS

The highest correlations are between scores on the language tests and on 'Polanyi' and 'Turbines'. Under both methods, the L/A group, linguistically the most competent, did better than the other 3 groups. One might, then, envisage scores as being made up of (a) a linguistic proficiency component; (b) a background knowledge component. On Engineering texts, the Engineers' proficiency component was smaller than the DAFE group, but their background knowledge component was greater, thus levelling scores. On Development texts, however, the DAFE group was superior on both components, hence their higher scores.

While linguistic proficiency would definitely seem to have been a factor, however, the explanation outlined above will not stand. It is true that in the DAFE group, the Lancaster Economists seem to have been more proficient; this was not true, however, of the Aston Development Admin/Finance group, who formed the larger part of the DAFE group as a whole. A calculation based on the Aston group alone indicated that there was still no difference between them and the Engineers on the Engineering texts, while they did better than the Engineers on the Development texts. Straight linguistic proficiency is not enough.

A rather more complex explanation involves linguistic proficiency, background knowledge, and the rank order of text difficulty. On the Gap-filling test, in both Study 1 and 2, 'Turbines' was the easiest text overall, with 'Electrolytes' second. These are, of course, the two Engineering texts. One could argue that on linguistically easy texts, groups at the same level of proficiency would score about the same. As the linguistic difficulty of the texts increased, scores based on proficiency would tend to drop away, and be replaced by scores achieved through bacground knowledge, if that was available. Thus on an 'easy' text, like 'Turbines', groups at the same level of linguistic proficiency, e.g. ENG, DAFE and S&M, would score roughly the same (the L/A group would score higher). On linguistically more difficult texts, e.g. 'Polanyi', ENG and S&M would be left behind. This does seem to explain some results. On 'Turbines', exact scoring, the scores were 30 (S&M), 32 (ENG), 33 (DAFE), and 43 /L/A).

However, neat though such an explanation may be, it cannot cover all the data. On Gap-filling, 'Electrolytes' was an easy text; on Short Answers, however, it was comparatiively difficult. Moreover, if it was linguistic simplicity that made 'Electrolytes' easy on Gap-filling, one would expect a high correlation between scores on that text and onELBA and ELTS. Examination of Table 6 shows that this was not the case. On both Short Answer and Gap-filling, correlations for this particular text are lower than for any other.

Conclusion: The evidence is that background knowledge has a significant effect. On the evidence of Study 2 scores, however, it does not seem to operate symetrically; Engineers were disadvantaged by being tested outside their speciality - the DAFE group were not. No satisfactory explanation has been found for this asymetry.

Study 3: The test-tasks used in Study 1 and 2 had certain advantages. Both Gap-filling and Short Answer tests are easy and quick to construct and administer, and reasonably easy to score. However, they are both open to charges of subjectivity. Pre-tests with suitable native-speakers, including subject specialists, and item-analyses of their responses, would help to make the tests less subjective. However, such procedures would, in effect, complicate tests whose major justification is their simplicity. Moreover, they would not help to improve the construct or face validity of such tasks used as proficiency tests.

We therefore decided to conduct a study using groups of students similar to those who took part in Studies 1 and 2, but replacing the

original test tasks with a module (M1) of the British Council ELTS test (1). Using part of the ELTS test had the following advantages:

1) the ELTS test is a sophisticated test, based on a complex, if somewhat non-empirical needs analysis (Munby), 1978).

2) it is the proficiency test used now by the British Council for assessing the language proficiency of potential scholarship students, and hence sat by a large proportion of tertiary-level foreign students wishing to study in Britain.

3) the study-skills module (M1) contains reading and study-skills material appropriate for the groups taking part in Studies 1 and 2.

Method: The tests used were (a) 3 subject-specific alternatives of the ELTS M1 Study Skills module, namely 'Social Science', 'Technology' and 'General Academic'; (b) the G1 (Reading) and G2 (Listening) modules of the ELTS test. The order of presentation was as follows:

1) G1, G2 and M1 (Social Science): at the start of the ESS course; used for placement purposes.

2) M1 (Technology): in week 6 of the 8 week course, week 2 of the 4 week.

3) M1 (General Academic): 2 weeks later; used as exit test.

Subjects: students belonged to the same general areas as before. Partly because of the constraints imposed by the experimental design, and the way the design was adapted to the administrative requirements of the ESS courses, the Engineers and the Science/Maths groups were combined on this occasion. This was justified by the evidence from Studies 1 and 2 that they are closely related groups. Only subjects who sat all three M1 tests were counted. Of these, 41 were DAFE, 35 S&M and E, and 41 L/A.

Results: Provisional results, based on band scores, are presented in Table 7. Since scores are still being processed, no significance values are included.

Study 3

Table 7

ELTS Scores

		G1	G2	M1 (Soc Sc)	M1 (Tech)	M1 (Gen Ac)	N
DAFE	X	5.02	5.74	4.62	5.12	5.82	41
	SD	1.049	1.215	1.254	1.654	1.105	
S/M&E	X	4.8	5.5	4.49	5.96	5.73	35
	SD	1.476	1.388	1.300	1.77	1.314	
L/A	X	5.88	6.35	5.22	5.45	6.37	41
	SD	0.914	1.2	1.280	1.219	0.873	
ENG	X	4.64	5.11	4.36	6.21	5.64	14
	SD	1.184	0.944	1.13	1.282	1.13	
S&M	X	4.9	5.76	4.57	5.79	5.79	21
	SD	1.663	1.586	1.417	2.047	1.445	

Discussion

If the hypothesis had been confirmed in all its details, each group would have scored best on the subject alternative most appropriate to its members, i.e. DAFE would have scored highest on 'Social Science', S/M and ENG on 'Technology' and L/A on 'General Academic'. Actual performance on 'Appropriate', 'Inappropriate' and 'Neutral' tests is shown in Table 8 (for L/A, there is no 'Neutral' test, although one might predict that they would do better on 'Social Science' than on 'Technology').

Table 8

Group performance on Appropriate/Inappropriate/Neutral Texts

	Approp	Inapprop	Neutral
DAFE	3	2	1
S/M&E	1	3	2
L/A	1	2/3	

With one exception, the table confirms the hypothesis and the assumptions underlying this part of the ELTS test, e.g. S/M&E did best on 'Technology', the most appropriate test, second best on 'General Academic', the 'neutral' text, and worst on 'Social Science', the 'inappropriate' test. The exception is the DAFE group, who did worst on the 'appropriate' test, i.e., 'Social Science'. Apart from this, even this group conformed to the hypothesis, doing better on the 'Neutral' test than on the 'inappropriate' one.

Two explanations are immediately available to account for this deviation from the hypothesis. The first is that the 'Social Science' module is simply more difficult than the others. Table 9 shows that all 3 groups found it hardest.

Table 9

M1 Rank Orders (1 = easiest)

	Soc Sc	Tech	Gen Acad
DAFE	3	2	1
S/M&E	3	1	2
L/A	3	2	1

While, however, this may be true, there is another explanation. The results may be confounded by the presence of a learning effect. With the partial exception of the S&M and E group, the subjects found the first test the most difficult, the second easier, and the third easiest of all. One may, of course, entertain some doubts about such a learning effect. Many of the students had attempted the M1 module before arriving on the course; they might have been expected to be familiar with the general form and requirements. Moreover, whatever one's faith in the efficacy of either Aston or Lancaster ESS courses, one may doubt such a learning effect on such a complex test over a period, for the 4-week students, of 2 weeks. The possibility remains, however, and precludes firm conclusions being drawn about the relative difficulty of the 'Social Science' alternative. We are left with the fact that the S&M and E group performed better than the other two groups on the 'Technology' module, and this cannot be explained by a learning effect. They did this, moreover, in spite of the fact that scores on G1, G2, and the other M1 alternatives suggest that the other two groups were in general more proficient. If one separates the Engineers from the S/M and E group (as on Table 7), their superiority on the 'Technology' alternative is even more striking.

Conclusions

The three studies reported above result in somewhat conflicting results. Study 1, without significance values, indicated that the overall hypothesis was correct in virtually all respects; Study 2

suggested that Engineers were disadvantaged on texts outside their area, whereas DAFE subjects were not; Study 3 showed the opposite.

Taken together, the studies do seem to confirm the importance of background discipline. Needless to say, further research is needed to resolve the contradictions.

Notes

1 The authors would like to thank the British Council for making these tests available for the experiment.

Bibliography

Alderson, J. C. and A. H. Urquhart. "The effect of student background discipline on comprehension: a pilot study". In Hughes, A. and D. Porter (eds). Current developments in language testing. London: Academic Press.

Gregory, D.P. 1972. Fuel Cells. London: Mills and Boon.

Hulme, B.G. "Development of off-shore turbine packages for power generation and mechanical drive." GEC Journal of Science and Technology 47.

Latham, A.J.H. 1978. The International economy and the under-developed world, 1864-1914. London: Croom Helm.

Munby, J. 1978. Communicative syllabus design. Cambridge: Cambridge University Press.

THE LEXICAL RICHNESS OF L2 WRITTEN PRODUCTIONS AND THE VALIDITY OF VOCABULARY TESTS

Pierre J.L. Arnaud (Université Lyon 2)

1 Introduction

Discrete-item tests of separate components of language, although still in use for practical reasons, have tended to fall out of fashion as language testing theory has begun to place more emphasis on validity. It seems, however, that vocabulary tests should be distinguished from the other categories.

In terms of content validity, even keeping in mind the often repeated caveat that a vocabulary is not a "bag of words", it is clear that words are by and large individual entities, and therefore that the lexicon lends itself well to the construction of discrete items. Furthermore, some types of vocabulary tests, such as those in which the testees are required to produce the name of illustrated objects or actions, correspond to the naming or labelling activity which is a frequent part of language behaviour in everyday life, and are immune from the accusation of eliciting only "language-like behaviour".

No reproach can be formulated against vocabulary tests on grounds of concurrent validity, either. In Ingram's (1968) words, the vocabulary test is "the nearest thing we have to a foolproof test, i.e. a constructor-proof test. The statistics, both item analysis and correlation figures, are nearly always good."

However, the ultimate proof of the validity of a test is external: do test scores, obtained in an artificial situation, reflect accurately the quality of language behaviour in real-life circumstances ? The purpose of the present research is to answer this question by comparing vocabulary test scores to indices of vocabulary richness of L2 productions.

2 Previous research

Vocabulary richness has been investigated intensively in lexicometric and stylostatistical research. This can provide useful methodological guidelines, in spite of the fact that the texts examined bear little resemblance to the productions of foreign-language learners.

By vocabulary richness, one usually understands the number of types (V) in a given length of text, i.e. in a given number of occurrences or tokens (N). Muller (1977:115) specifically excludes any other measure, such as the degree of rareness of the vocabulary, which is not a "fact of structure", but a "fact of content". The comparison of texts of different lengths in terms of lexical richness poses a number of problems, as new words occur in decreasing numbers as a text unfolds itself. This can be explained by means of the "urn model":

the lexicon of a writer can be compared to an urn, from which a lexeme/type is drawn each time a token is produced in discourse; the type is then placed back into the urn, from which it can be drawn again. Supposing the lexicon of a writer is not infinite, there remain fewer and fewer lexemes that have not been drawn before. An index such as V/N is therefore unstable. Other indices for the comparison of texts of different lengths have been experimented with, but they have proved disappointing (Ménard 1972:105f.; Muller 1977:117; Dugast 1978). In order to make legitimate comparisons, it is therefore necessary to shorten the longer text, which can be done by actually deleting tokens or, alternatively, by using Muller's method of simulated draws based on the binomial law: knowing the distribution of frequencies for a given length of text, it is possible to calculate the theoretical V' for an inferior length (Muller 1977:101f.,127f.). It should be noted, however, that the confidence interval of V' becomes wider as the original text is more radically shortened.

Ménard (1972,1978) has investigated the relationship between lexical richness, as defined above, and the presence of rare words in a text. After comparing the effects of various empirical norms of rareness, he concludes that too restrictive a definition (such as absence from a dictionary) leads to a Poisson distribution of rare words, thus making statistical treatment difficult. When more liberal norms are applied, highly significant rank correlations appear between lexical richness and the proportion of rare words in V (1).

Less attention has been devoted to lexical richness by applied linguists and the author has found only two publications on the subject: Linnarud (1975) and Mendelsohn (1981). Linnarud, in a pilot study, examines a corpus of 36 essays written in English by Swedish students; in addition, three native speakers, enrolled in the same course, wrote an essay on the same subject. As the author herself points out, the choice of the subject ("Sir, I protest...") was unfortunate as it led to very divergent responses and thus to poor control of vocabulary content. The essays varied in length from 124 to 573 words. Two main variables were taken into account: lexical density and lexical variation. Lexical density is the percentage of lexical words in the text: LD = $(N_{lex} \times 100) / N$. Linnarud mentions difficulties due to the lack of firm criteria for distinguishing lexical words from grammatical ones, and the need to make carefully weighed decisions in some cases. As some essays had a high LD, but contained a poor, repetitious vocabulary, lexical variation was also measured, with the following formula: LV = $(V_{lex} \times 100) / N_{lex}$. Measures were collected once with errors deleted and once with errors retained; only the first series is reported here. The lexical density statistics (minimum: 30.86; mean: 38.51; maximum: 47.15) fall below the values observed by Ure (1971) on L1 texts. The LD of the essays written by native speakers is higher. As the different lengths of the essays made their comparison in terms of lexical variation illegitimate, Linnarud divided them into four groups according to length, and within each group classified the essays into three lexical variation categories: here, too, the native speakers scored better; their essays also contained a higher percentage of hapax. Finally, the essays were classified into three groups according to their global value, and groups were compared in terms of mean length: the "poor" essays are the shortest, but the "average", and not the

"good" ones, are the longest. However, the simple statistical treatments used in this pilot research provide suggestions more than definitive results.

Mendelsohn (1981) studied the oral productions of 42 subjects, native speakers of English and learners. He measured three variables that, according to him, constitute lexical richness. Lexical density was established by means of a formula slightly different from Linnarud's, but which would yield the same rank orders: LD = $(N_{lex} \times 100)/N_{gr}$. Like Linnarud, Mendelsohn mentions the difficulties due to the lexical/grammatical distinction, and recommends keeping notes of decisions made in ambiguous cases. Lexical variation was also measured using a slightly different method: LV = $(V_{lex} \times 100) / N$. A "semantic variation" measure was introduced by Mendelsohn: when students are required to speak on a specific subject, a certain number of "universes of discourse" can be expected to appear in their productions, and the number of types belonging to them can be counted. Mendelsohn found that this was the most discriminating of the three measures of lexical richness. In addition, he measured the percentage of errors. As LD and LV yielded a rank-correlation coefficient of .78 and error percentage and lexical richness one of only .395, the author concludes that lexical richness and errors represent different dimensions of lexical performance, and that both should be taken into account for evaluation purposes. In most situations, however, it seems that errors are the only criterion applied: observation of a group of 35 evaluators working on the corpus showed that subjective judgments were mainly based on them, and that lexical richness had little influence. Indeed, another group, required to classify subjectively a set of transcripts for lexical richness, performed poorly on the task.

Other studies dealing with the assessment of written productions of L2 learners have yielded interesting results concerning the role of vocabulary in the formation of judgments. Perkins (1981), comparing three methods of essay evaluation: global subjective, analytic subjective (i.e. a group of judgments on criteria such as relevance, fluency, grammar, mechanics, vocabulary), and objective (number of errors, number of words per T-unit) found that, among the five subjective criteria, vocabulary had the lowest correlations with the other variables and the subjective global score. On the contrary, Mullen (1980), working on a corpus of 117 texts assessed by four judges globally and analytically along four dimensions, structure, organization of ideas, quantity of text produced and vocabulary, found that vocabulary was best correlated with the global score.

Finally, the relationship between essay length and quality has been investigated. Linnarud's (1975) results have already been quoted. Larsen-Freeman and Strom (1977) and Neumann (1977), as quoted by Perkins (1980), found a significant correlation between essay length and overall quality, whereas Perkins himself did not observe such a relationship. Again, Mullen (1980) reports a high correlation between quantity of text and global score, but this quantity itself was estimated subjectively.

3 Method

3.1 Administration

The subjects were first-year specialist students of English at Lyon 2 University. In order to reach a sufficient population, the experiment was carried out during normal teaching hours. The vocabulary test was administered first; it was a 26-item productive test (translation from French into English) with a reliability (K.R./Horst) of .89 (Arnaud 1981). On the following week, an hour was devoted to the essay writing. To ensure motivation, global scores were taken into account for continuous assessment. The subject was chosen so as to give every student something to say:

"What do you suggest to improve secondary education in France ?"

Data treatment possibilities led to the retention of a random sample of 100 test/essay pairs. In addition, four sophomores from Dartmouth College, then at Lyon 2 on a regular programme, wrote an essay on the following, symmetrical subject:

"What do you suggest to improve high school education in the U.S.A. ?"

Although this control group is too small to allow for statistical comparisons, and in spite of the fact that the local students were hardly comparable to those in academic terms, a useful reference is thus provided.

3.2 General considerations

The essays were first reviewed subjectively. An essay can be seen as the conjunction of a number of ideas and their connections on one hand and the linguistic means to express them on the other hand. There exists a much larger degree of independence between the two in the case of L2 learners than in the case of native speakers, and on examining a poor essay it is not always easy to determine what is due to a paucity of ideas or to low language proficiency. Furthermore, when one tries to concentrate on the purely linguistic aspects, clear-cut components seldom appear prominently: for instance, one would hardly ever come across an essay in which the sentence grammar was impeccable and the text grammar disastrous. The following sample is typical of this kind of problem:

"We should find other way for holiday. The summer holidays are too long. During often three months, the pupils forget what they have learnt in the year. It should be better to have more holiday but shorter holiday and to finish the school at 4 o'clock and not at 6 o'clock."

Such repetition might be attributed to lack of vocabulary or ignorance of cohesive devices, but the fragment happens to be strikingly similar to the following (oral, L1) passage discussed by de Beaugrande and Dressler (1981:54):

"There's water through many homes. I would say all of them have water in them. It's just completely under water."

The authors suggest that this discourse is due to lack of planning time and rapid loss of the surface text, two phenomena which obviously play a role when students at this proficiency level have to write English texts in a limited time. An essay including many repetitions like the one quoted from above would result in a very low lexical

variation score, which would however to a great extent be the effect of non-lexical causes. This should serve as a reminder that total isolation of one component is probably illusory and that one should not place excessive trust in objective measures of written text variables.

A number of facts about lexical quality do appear clearly, however. As impressions are produced essentially by lexical words, non-lexical tokens can be left aside, as in Linnarud's (1975) and Mendelsohn's (1981) studies. Some of the essays strike the reader as containing a very limited number of lexical items repeated interminably; this is so widespread a phenomenon, and one that lowers the overall quality of an essay so much that it must be taken into account in any definition of L2 lexical richness. As Mendelsohn (1981) had found that lexical density and lexical variation are highly correlated, it was decided to determine lexical variation through the following formula:

$$LV = \frac{V_{lex}}{N}$$

In other cases, attention is attracted to the vague, general vocabulary used by the writer. For instance, the only way some students were able to express value judgments was literally to use good and not good. What is involved here is the degree of rareness of the vocabulary and it was therefore decided to assess it.

Finally, the number of lexical errors is sometimes so high as to hinder communication. Naturally, the author is well aware that, as a non-native speaker of English, he may be more intolerant of errors than a native speaker: James (1977) found that it is the non-natives who are less tolerant of lexical errors; Hughes and Lascaratou (1982) have shown that non-natives react to the fundamental nature of the rules that are transgressed, whereas native speakers are sensitive to the way communication is affected. Mendelsohn (1981), as seen above, has brought to light the importance of errors in the subjective evaluation of lexical quality. The present research therefore differs from classical stylistic studies of L1 texts in that it includes the measurement of errors and vocabulary rareness.

3.3 Data treatment

3.3.1 Editing norms

The texts had to be treated on a small computer (2) with programs written by the author, and this required great simplicity in the operations, which made the following editing norms necessary.

Proper nouns (except country names) and numbers were deleted.

Compound words: after their existence as items was checked in a dictionary (Hornby 1974), they were typed hyphenated and thus treated as units by the program.

Verb-particle constructions were typed separated or hyphenated according to the case:

He turned down the street.
He turned-down the offer.

As syntactic errors were not of direct relevance to this research, they were typed verbatim. In most cases, this has no influence on the lexical richness indices:

There is too much pupils in the schools,

but N is sometimes affected:

The philosophy is not properly taught.

This is another proof of the impossibility to isolate lexical aspects of writing entirely.

Morphological errors were corrected when it was clear that knowledge of the lexeme was not affected (but see below).

Lexical errors: although we are not concerned here with error analysis, such regularities occurred that a brief survey is of interest; note the prevailing influence of interference.

minor spelling mistakes: personnal, teatcher
major spelling mistakes: scholl
derivation mistakes: to comparate, he successed
faux-amis (deceptive cognates): They should be prevented that it is difficult
interference from another language on the curriculum: to spare money
confusion between two lexemes: The teachers learn them maths.

Although the gravity of these types of errors is variable, an all-or-nothing system was applied: any lexically erroneous form was replaced with an error code, and was counted in N but not in V.

Borderline cases: some morphological errors indicate that the corresponding lexeme is not stored correctly: errors on irregular verbs were treated as lexical. Similarly, the preposition in the following example was treated as a lexical mistake, as one cannot consider that a lexeme is known if the accompanying structures are not mastered:

They listen at the teacher.

Prepositions were otherwise treated as grammatical items.

Problems due to the subject: the lack of correspondence between the American, British and French educational systems caused a number of problems. When a student used a term from an Anglo-saxon context to designate a French reality, this was not considered as a mistake, but the word was ignored. Furthermore, it was often difficult to determine whether a particular use was legitimate or constituted an interference error, as the context was ambiguous: for instance, it was seldom clear whether minister stood for ministère or ministre, and if class was intended to mean class or classroom (both classe in French). The benefit of the doubt was systematically accorded, but this introduced a further lack of precision in the measurement.

3.3.2 Lexical vs. grammatical words

For automatic treatment purposes, it was out of the question to have to make a decision every time an ambiguous form turned up; tokens like have or do were therefore systematically treated as grammatical. As a filter for non-lexical occurrences, the program included a dictionary of grammatical forms to which each token was compared in turn. As treatment time grew exponentially with the size of this dictionary, it had to be limited to approximately 100 items. These are the first grammatical items in the rank-order list of the Hofland and Johansson (1982) frequency count. The classical closed class/open class distinction was therefore not respected, as many closed-class members were left outside the dictionary, but this was not really a weakness, for few of the grammatical items present in the corpus passed through the filter, given the proficiency level of the subjects; furthermore, knowledge of a rare grammatical word cannot be considered as independent from lexical knowledge.

3.3.3 Lemmatization

Lemmatization was carried out manually on the print-outs. Plurals were grouped together with the singulars, conjugated forms with the infinitive, comparatives and superlatives with the positive degree. Only one lemma was counted for forms belonging to two categories (book's, booked).

3.3.4 Lengths and indices

The distribution of essay lengths is represented in Figure 1. Variation is considerable.

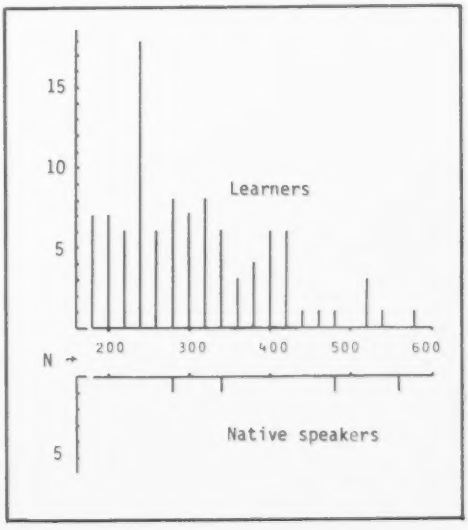


Figure 1

distribution of N (essay lengths)

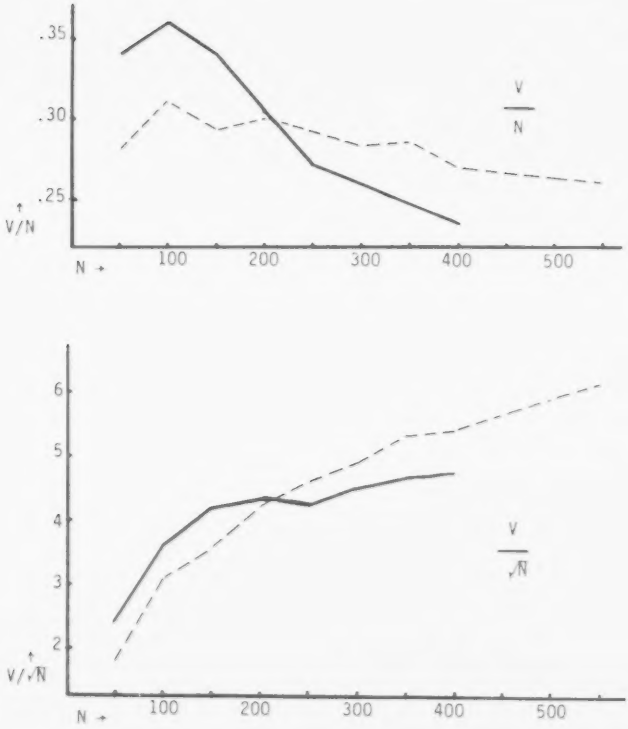
M = 313

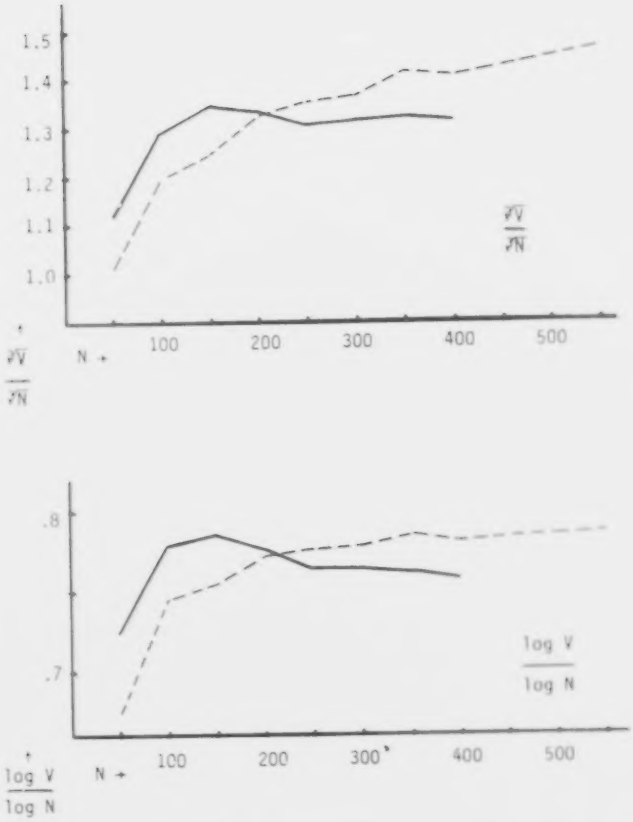
SV = 91

(N_{Ss}: 100 + 4)

Figure 2

Evolution of V/N and other indices: successive 50-word fragments of two essays





The graphs of Figure 2 prove empirically that V_{lex}/N and indices derived from it are unsuitable for the establishment of lexical variation scores on L2 texts in this length range. It was therefore necessary to bring down all the essays to the length of the shortest one, N = 180. Use of Muller's (1977) method of simulated draws would have been exceedingly time-consuming, so token deletion was chosen instead. It was impossible to retain, for instance, the first 180 tokens of each essay as this would not have ensured adequate sampling. A series of instructions was therefore added to the program for the random selection of 180 tokens. The lexical variation score hence-

forth corresponds simply to V_{lex} and the error score (E) to the number of lexical errors in the sample.

3.3.5 Norm of rareness

It would have been theoretically possible to calculate the average frequency of the types in each essay, using a frequency list, but such a task would have proved unsurmountable practically. An empirical solution like Ménard's (1978) was therefore retained, and the norm was chosen as wide as possible. The official French Ministry of Education vocabulary list for the classe de troisième (Elaboration 1976) was used. It contains 1 522 lexical items that are supposed to be known of all students by the time they reach the lycée level. A type was therefore considered as rare if it was not on the list. The following formula was used for establishing the rareness scores:

$$R = \frac{V_{rare}}{V_{lex}}$$

3.3.6 Reliabilities

As text shortening introduces a margin of uncertainty, the reliabilities of V, R and E were checked. The program was run four times over a sample of 20 essays, each time extracting a different set of 180 tokens. Product-moment coefficients of correlation were computed between the values yielded by runs I and II, and III and IV respectively. The results of the first and last two runs were then averaged, and the correlations between the two sets of means calculated. The figures therefore correspond to reliability coefficients for values obtained after two runs of the program, which was the case for the remaining 80 (+4) essays (Table 1).

Table 1: reliability coefficients of lexical richness variables after text shortening; bottom line: r's for mean values after two runs of the program (20 texts)

V				R				E			
I	II	III	IV	I	II	III	IV	I	II	III	IV
.88		.73		.86		.90		.78		.97	
.87				.95				.92			

3.3.7 Distributions

The distributions of V, R and E are reproduced in Figures 3, 4 and 5. Interestingly, in terms of lexical variation and rareness of vocabulary, some French students performed as well as their American counterparts. The error scores of two of the latter are due to spelling mistakes.

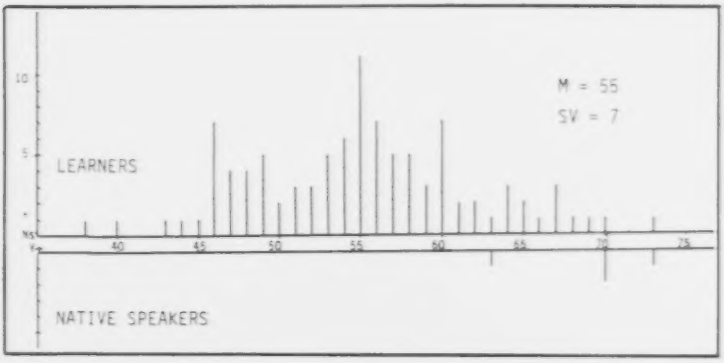


Figure 3: distribution of V

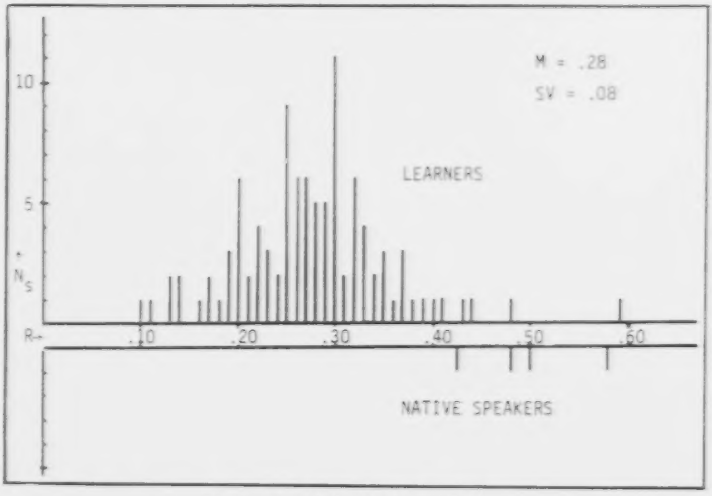


Figure 4: distribution of R

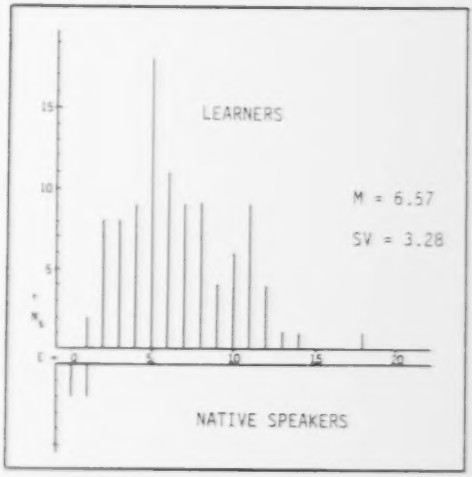


Figure 5: distribution of E

4 Results and discussion

It is not unreasonable to think that the more proficient students are able to write longer texts in a limited time. Subjective examination, however, had shown that some of the longest essays were among the most repetitive and unconnected. This impression is confirmed, as far as vocabulary richness is concerned, by the correlation coefficients in Table 2, none of which is significant. In view of the contradictory results of previous research (see section 2), it seems that no generalizations should be made in this respect.

Table 2: correlations between essay length and the three lexical richness variables as established on 180-token samples

r_{NV}	=	.06
r_{NR}	=	.10
r_{NE}	=	.05

Product-moment coefficients of correlation between test scores (T) and the three lexical richness variables are reproduced in Table 3. There obtain significant, but low correlations between V and R and V and E (3); R appears to be uncorrelated to T and E. However, it would be unwise to conclude that R measures a different dimension of lexical richness before replication studies can be carried out.

Table 3: correlations of test scores (T) and lexical richness variables (N_s = 100); * : p<.025; ** : p<.01

	T	V	R	E
T		.36**	.09	-.21*
V			.27**	-.24**
R				.09

At this stage, a few points should be remembered. First, a test score is the result of a concentrated measurement situation, which is not the case of an essay. Secondly, a substantial amount of uncertainty was introduced successively by editing norms, decisions as to errors and text shortening; absolute isolation of the lexical component was furthermore impossible. Finally, the students, all native speakers of French with fresh baccalauréats, constituted a very homogeneous group and this obviously lowered statistics (see Raatz and Klein-Braley 1981). In view of these considerations, the low, but significant correlations between T and V and T and E make it possible to conclude that discrete-item vocabulary tests are valid.

5 Suggestions for further research

One of the questions which it has not been possible to answer in this study is that of the concurrent validity of the lexical richness measures. The only way to do this would be to analyse two sets of essays written by the same subjects, which was administratively beyond the author's reach.

For further research, use of the following empirical index is suggested:

$$LR = V_{lex} + V_{rare} - 2E$$

(on text samples of the same length). Rare types and errors are given double weight. This index was found to separate native speakers from learners very clearly.

Notes

1 Ménard, however, does not prove that an index like V_{rare} / V is independent from the size of V.

2 The author wishes to thank his colleagues of the Service Informatique, Université Lyon 2 for providing unlimited access to the computer and indispensable advice.

3 See Note 1 for a possible dependence between R and V. Furthermore, as error words are at the same time counted in E and ignored in the establishment of V, the two measures are not entirely independent.

Bibliography

Arnaud, P.J.L. 1981. Comparaison pratique de cinq types de tests de vocabulaire anglais et recherche sur la nature de la compétence en vocabulaire. ms. (to be presented at the 7th World Congress of Applied Linguistics, Brussels, Aug. 5-10, 1984)

Beaugrande, R.de, and W. Dressler. 1981. Introduction to text linguistics. London: Longman

Culhane, T., Klein-Braley, C., and D.K. Stevenson (eds). 1981. Practice and problems in language testing. University of Essex, Department of Language and Linguistics, Occasional Papers 26

Dugast, D. 1978. "Sur quoi se fonde la notion d'étendue théorique du vocabulaire?". Le Français Moderne 46: 25-32

Elaboration d'un programme lexical pour l'enseignement de l'anglais dans le premier cycle. 1976. Paris: I.N.R.D.P., Recherches Pédagogiques 84

Hofland, K., and S. Johansson. 1982. Word-frequencies in British and American English. Bergen: The Norwegian Computing Centre for the Humanities

Hornby, A.S. 1974. Oxford advanced learner's dictionary of current English. London: O.U.P.

Hughes, A., and C.Lascaratou. 1982. "Competing criteria for error gravity". English Language Teaching Journal 36: 175-181

Ingram, E. 1968. "Attainment and diagnostic testing". In Davies, A. (ed). 1968. Language testing symposium. London: O.U.P. 70-97

James, C. 1977. "Judgments of error gravities". English Language Teaching Journal 31: 116-124

Larsen-Freeman, D., and V. Strom. 1977. "The search for a second-language acquisition index of development". Workpapers in TESL 11: 35-43

Linnarud, M. 1975. Lexis in free production: An analysis of the lexical texture of Swedish students' written work. University of Lund, Department of English: Swedish-English Contrastive Studies, report n°6

Menard, N. 1972. Mesure de la richesse lexicale, Méthodologie et vérifications expérimentales. Thèse de troisième cycle, Université Strasbourg 2

- . 1978. "Richesse lexicale et mots rares". Le Français Moderne 46: 33-43

Mendelsohn, D.J. 1981. We should assess lexical richness, not only lexical error. Paper presented at the 1981 TESOL convention, Detroit

Mullen, K.A. 1980. "Evaluating Writing Proficiency in ESL". In Oller, J.W., and K. Perkins (eds). 1980. Research in language testing. Rowley: Newbury House. 160-170

Muller, C. 1977. Principes et méthodes de statistique lexicale. Paris: Hachette

Neumann, R. 1977. An attempt to define through error analysis the intermediate level at UCLA. unpublished thesis, UCLA

Perkins, K. 1980. "Using objective methods of attained proficiency to discriminate among holistic evaluations". TESOL Quarterly 14: 61-69

- . 1981. "The test of ability to subordinate: Predictive and concurrent validity for attained ESL composition". In Culhane & al. (eds). 1981. 104-112

Raatz, U., and C.Klein-Braley. 1981. "The C-test - A modification of the cloze procedure". In Culhane & al. (eds). 1981. 113-138

Ure, J. 1971. "Lexical density and register differentiation". In Perren, G.E., and J.L.M. Trim (eds). 1971. Applications of linguistics. Cambridge: C.U.P. 443-452

THE CONSTRUCTION OF A MODULAR THEMATIC TASK BOUND ENGLISH LANGUAGE ACHIEVEMENT TEST BATTERY FOR KUWAIT

M.I. Baghdady (The Language Centre, University of Exeter)

ABSTRACT

The purpose of this paper is to present a research progress report on the developmental stages, the construction process, the administration procedures and assessment scales of a Modular Thematic Task Bound English Language Achievement Test Battery for Kuwait.

An assessment of the need for such a battery will be attempted as well as a description of its modules, themes, tasks and the criteria used for the sampling of these tasks from the syllabus and course objectives.

Samples of the battery tasks and the administration and scoring manual accompanying it will be given as well as the task analysis and componential inventory. The pilot scheme and validation process planned for this battery will also be reported on.

The Research and its Aims

The construction of this battery is part of a Ph.D. research programme conducted by the researcher and has been in progress at the Language Centre of Exeter University since October, 1981. The aims of this research are:-

1. an investigation into the types of tests appropriate to assess the English language achievement of students of English as a first foreign language studying the "Crescent English Course", (ELTA, O.U.P., 1977-) at the mid-secondary school level of general education in Kuwait,

2. the production of relevant and appropriate guidelines for future test construction by teachers and inspectors at all levels of general education in Kuwait, and

3. on the basis of an assessment of the currently practised testing policies, techniques and systems used there, the research broadly aims as well at initiating a basic overhaul of the testing and examinations approach within a framework of integrated educational reform of curricula in general, and ELT in particular.

Recent ELT Development in Kuwait and the "Crescent English Course"

As a phase of the educational development movement in Kuwait, the Ministry of Education there decided in 1976 on a redefinition of the ELT objectives in the country. That decision was a result of:

(a) A 3 year research into the outcomes of the then used structurally-based course. That research highlighted a considerable dissatisfaction with that course on the parts of educationists, parents and learners in terms of the levels of achievement in English at that time which were very low.

(b) A recognition of the need for a communicative approach to ELT in which the notional and functional values of language use should be focused as the only alternative for the ELT syllabus to meet the requirements of societal movement and the learner's needs.

Hence, the ELT objectives in Kuwait were redefined through a research process and democratic procedures, specifically emphasizing:

(a) the integration of language skills in a classroom context of activity and variety

(b) the placement of the learner at the centre of the educational process

(c) the use of authentic language teaching/learning materials as much as possible within the constraints of the classroom situation

(d) the integration of multi-media with study-materials

(e) the development of an attitude to learners' errors acknowledging their transitional competence

(f) functionalism in the design and assignment of teaching/learning experiences

(g) auto-learning or self-study strategies.

The Crescent English Course (henceforth CEC) by ELTA/O.U.P. was introduced to Kuwait initially as a project for the development of ELT, which was basically concordant with the new orientations in the educational scene at that time. Later, the CEC materials were introduced to Kuwait as tentative representations of an English language teaching syllabus for the Arab World. As O'Neill and Snow (1977) stated in the introduction to these materials, they are based on:

> "the current appreciation of the communicative nature of language and on an understanding of the implications this has for teaching and learning."
>
> O'Neill, T. & P. Snow, 1977
> "Crescent English Course: Teacher's Book 1"
> p. 1, ELTA - O.U.P., 1977, 1978, 1982.

The concerned Institutions of the Ministry of Education, Kuwait spent four years analysing and experimenting with the CEC materials to decide on their suitability and relevance to the ELT objectives for general education in the country. *The writer reported on this process in detail elsewhere. Finally, the course has been fully adopted as the official ELT course in the general education curriculum of the State, where English is taught starting from the beginning of the Intermediate stage, i.e. the 5th year of schooling, to the end of the secondary stage, i.e. the 12th and final year of schooling. The school ladder in Kuwait consists of twelve years divided equally into three stages, Primary, Intermediate and Secondary and English is a general requirement subject for two thirds of this time. Throughout the eight years English is given an average of five lessons a week. The CEC is a series of eight levels, each for one of the eight years of the ELT programme. The course components for each level comprise:

A Teacher's Book
A Pupil's Book
A Pupil's Reference book. (Applicable to the secondary stage)
A Pupil's Workbook
Recorded Materials: Class and Pupil's Cassettes
O.H.P. Transparencies
Supplementary Readers

The materials are arranged for each level in Units sub-divided into steps, the aims of which are stated at the beginning of each one. The course is claimed to be activity-based emphasizing communication, socialization, individualization and enjoyment.

The Need for the Battery

It may be worth mentioning here that a paradoxical situation currently exists in Kuwait where learners are taught to learn English by a communicatively-based approach whilst they are still being tested to a considerable extent as they were being tested in the light of the structurally-based approach. (1) It is not a happy situation in which pupils, albeit unintentionally, are unfairly tested, and teachers are at a loss when it comes to communicatively-based language tests. The missing link in the whole process of ELT syllabus development there is testing. Besides, testing and testing policies in Kuwait have always been a problem

* See Baghdady, M.I., (1982, a,b,c) in Bibliography

calling for solution. Apart from some ad hoc specimen tests provided by the course writers, and similar ad hoc tests produced by some teachers and some attempts on the part of Inspectors at setting up tests to match the new approach, the question of testing in the light of the new Syllabus is yet without a satisfactory answer.

The Rationale of its construction

The battery is meant to be 1. relevant to its purpose as an achievement test with the course objectives as the frame of reference, 2. reflective of learners' state of learning, 3. revisory of syllabus through its expected feedback on the course outcome, 4. referential in the sense that performance elicited is contextualized, and 5. rateable, i.e. profiling achievement.

It is also based on level 6 of the CEC (i.e. the second year secondary which is the mid-secondary school level) as a prototype for the other levels in the series, demonstrating a methodological approach that is applicable to the other levels in the series pre- and post-level 6. Thus, the current battery which is trying to assess level 6's input could be considered a model for testing which, with slight modifications, could be adaptable to lower or higher levels. Performance and accountability are targets for this battery in the sense that it is intended to gauge the outcomes of a specific course of study, a defined set of experiences in terms of the learners' attainment of a certain performance level in language (i.e. a course-oriented achievement test). It is also designed hopefully to provide a fair scale to base decisions of learners' promotion from one level into another depending on the learner's achievement of the syllabus/course objectives. It is also hoped that this battery provides qualitative and quantitative feedback data for the evaluation, remediation and further development of the syllabus and the quality of training in particular. Most important is the endeavour to construct this battery of language tasks which can be tested, are fairly objectivized without resorting to the multiple choice format whilst providing considerable elicitation of language productivity on the part of the learner/testee at the specified level. Because it is communication oriented it is production based as well, since communication in language is, in a sense, synonymous to language production. This is why the central aim of the battery is to go beyond the lower order level of mere recognition to the elicitation of varied samples of the learners' purposeful production of the language in a clearly detailed context where the learner/testee has got a task to perform. It is also meant to be personalizing in the sense that it addresses the testee in person stimulating him/her to play a role in language activities bringing up his/her personality in performance. Hence, instructions are worded as friendly requests. Local and national environment is evident throughout the battery and the various tasks and situations are true to the testee's life and age so that the battery will appeal to the subjects.

The Construction Process

Because of the test specificity (i.e. an achievement test of a specific level of a specific course for specific students), the sampling of tasks is necessarily from what is there in the course with which the testees are familiar. The syllabus and course objectives are the criteria and "record of the outcome". (Cronbach, 1964: 108). Looking through levels 1-6 of the CEC and getting the type of model for what is being taught, the content of the course in that sense dictated the kinds of tasks required by the test. So, the types of activities and tasks assigned by the course were listed for a complete framework then concentrating on level 6 for a detailed componential inventory in terms of tasks and activities. After preparing a basic statement about what students should be able to do at the end of level 6, sampling of these tasks was done implementing the following sampling criteria:-

(a) Surveyability: i.e. representative of broad abilities.

(b) Clearness: i.e. freedom from tricks and ambiguities.

(c) Proportionality: i.e. attempting a balance between and among tasks.

(d) Conciseness: i.e. spatio-temporally economical.

(e) Pragmaticality: i.e. true to life, useful and usable authenticated tasks within a framework of concessions to peculiarities of classroom situation.

The battery then was arranged in four modules under the general theme: "In-Kuwait English" indicating that the language situations in the test are types of the situations where English is used in the Kuwaiti society. Of course the situations sampled in the battery are not exhaustive but representative. If some of the situations devised for the tasks seem a bit contrived, it is certainly the result of the restrictions imposed by the limitations of the classroom situation.

Each module deals with a cluster of integrated language tasks interlinked by a modular theme which is consistently kept and observed all through the module. By module is meant an independent and self-contained unit of integrated constructional components.

The battery was discussed in manuscript in detail by a committee at the language centre established for this particular purpose and comprising linguists, educationists and expert testers. One of the course writers was invited to the committee meeting as a member representing the point of view of the course writers. After detailed discussions of the battery and its pilot scheme the committee approved the battery in its present form with an imprimatur. The present writer's indebtedness to this committee (2) is great for their very useful remarks, but any mistakes in the battery are his own responsibility.

The nature of the format

The battery comprises four modules of integrated language activities thematically linked, hence it is called "A Modular Thematic Task-Bound Battery". (See Table 1 below).

TABLE 1. THE BATTERY COMPONENTS GENERAL LAY-OUT

Module No.	Theme	Integrated Activities	Performative Aims	No.of Tasks	Time Allowed Hrs.Mins.	Top Level Achievement Scores
1 A,B	Kuwait	R ⟶ W	Reading (for gist, for specific information and for detail) leading to writing.	4	1 30	100
2	Emergencies	W. cued by varied sources	Transcoding information from various written sources into other various types of writing.	4	1 20	100
3	Social Events/ Links	T. cued by varied sources	Talking cued by Reading, guided by writing and/or Listening. (A telephone message - Language-Transfer).	4	- 45	100
4 A,B	Points of View	L ⟶ W	Listening (for gist, for specific information and for detail) leading to writing.	4	- 30	100

Code: R = Reading W = Writing T = Talking L = Listening

COMPONENTIAL DESCRIPTION OF EACH MODULE IN TERMS OF TASKS

It may be worth pointing out here that M.1 deals with various types of reading tasks, i.e. for gist, for specific information, for detail, and for reconstruction of a written discourse (cloze). (See Table 2 below).

TABLE 2. COMPONENTS OF M.1

Module and Theme	Task No.	Component Tasks Task Type & Aim	Time Hrs.Mins	Alternative/ Parallel Forms Group A	Group B
1 Kuwait A,B	1	Form filling, (though irrelevant to this module, and definitely relevant to module 2, yet by necessity it should appear at the very beginning of the test according to the ordering/arrangement of the battery to provide the testees' personal data	- 10	Same form with data arrange-ment different from Group B	Same form with data arrange-ment different from Group A
	2	Scanning a text for information (i.e. selective reading) and information transfer (inter-pretation of tabulation) and elicitation and writing in note form	- 25	Same text as B but question requisites are diff-erent from though equal to B	Same text as A but question requisites are diff-erent from though equal to A
	3	Cloze - procedures Reconstruction of a text with every 10th word deleted. (Knowledge of language hollistic-ally and of the flow of discourse).	- 45	Uniform	Uniform
	4	Skimming a text for gist and giving a title (i.e. - com-pression of information in a nutshell).	- 05	Different (from B) though equal text with identical question requisite	Different (from A) though equal text with identical question requisite

M.2. deals with various types of writing tasks, i.e. extended writing, compressed writing (telegraphese), writing to report events/accidents, to express feelings, or reactions to events and accidents, writing to warn and writing to transform information from one language into another ... etc. (see Table 3 below).

TABLE 3. COMPONENTS OF M.2

Module and Theme	Task No.	Component Tasks Task Type & Aim	Time Hrs	 Mins	Alternative/ Parallel forms Group A	 Group B
2 Emergencies	1	- Writing a letter/reporting to a newspaper. - Transcoding information from note into detail. - Reading and interpretation of forms leading to description of events/accidents in detailed writing		- 30	UNIFORM MODULAR FORMAT SINCE THE ACTIVITIES REQUIRED OF TESTEES ARE OF INDIVIDUAL PRODUCTIVE NATURE	
	2	- Writing telegraphese - Transcoding information from notes into a telegram form by compression		- 15		
	3	- Guided writing cued by reading. - Generation of information from given information. (Generation of warning from advice). - Writing to warn others against possible dangers - Inference		- 20		
	4	- Language Information transfer - (Translating)		- 15		

M.3 deals with talking cued by varied sources contextualized in a long-distance telephone conversation where the testee plays the part of the conversation initiator while the part played by the second interlocutor is provided on the tape and in the test booklet. (See Table 4 below).

TABLE 4. COMPONENTS OF M.3

Module and Theme	Task No.	Component Tasks Task Type & Aim	Time Hrs Mins	Alternative/ Parallel forms Group A Group B
3 Social Events	1	- Making notes in preparation for the focal task	- 15	UNIFORM MODULAR FORMAT SINCE ACTIVITIES ARE OF AN INDIVIDUAL PRODUCTIVE NATURE
	2	- Filling in a missing part on a dialogue format in preparation for the main task	- 15	
	3	- A telephone call (transmitting a telephone message)	- 1½	
	4	- Language information transfer (integration of L1 and L2 in an interpretation context).	- 10 (preparation) and - 1 (actual performance)	

The setting for this test is the language laboratory for the following reasons:-

1. It is neither practical nor psychologically healthy to test natural communication in a classroom test situation by establishing an interviewing committee. It is thought to have a psychological pressure on the subject which might block the ability to speak, at least with introverts or shy personalities. As Cronbach (Cronbach, 1964:30) writes, "Any test performance depends on both ability and personality."

2. The telephone conversation simulated in the language laboratory is a maximum approximation to the situation in real life while the pressure of the testing situation on the testee is minimized. The headphone and microphones bear considerable resemblance to the telephone set in terms of function and hypothetical use.

3. Conducting the test in the language laboratory allows for 30 subjects to perform their tasks in the test at the same time, which is an aspect of practicality.

4. The difficulty level of real life communicative situation

with "unpredictability" (Morrow, 1979) as one of its features, is reduced by providing the testee with the second part of the conversation to suit the target level which is not an ultimate proficiency one.

Module 4 deals with various levels of listening (i.e. for gist, for specific information and for detail) testing the understanding of spoken English. (See Table 5 below).

TABLE 5. COMPONENTS OF M.4

Module and Theme	Task No.	Component Tasks: Task Type & Aim	Time: Hrs	Time: Mins	Alternative/Parallel forms: Group A	Alternative/Parallel forms: Group B
4 Points of View A,B	1, 2 & 3	Based on listening answering (in writing) three pre-listening questions on gist, specific information and detail successively.	-	18	Same listening passage as for B with different though equal focusing	Same listening passage as for A with different though equal focusing
	4	On the basis of listening - expressing personal opinion and reasoning in an extended piece of writing	-	12		

This module is conducted in the language laboratory as well. A contextualized taped conversation between a father and a daughter on the advantages and disadvantages of T.V. in our modern life is played. The purpose of the listening is set for the subjects by the pre-listening questions in the test booklet. Each question focuses attention on a certain aspect of the dialogue and is designed to provide information about testees' abilities in a specific listening type.

Modules 1 and 4 were designed in two alternative or parallel forms differently arranged from but equal to each other in terms of types and levels of tasks. This was done for the following reasons:

1. security purposes, i.e. to prevent cheating or copying, and
2. future reliability studies, i.e. parallel test correlation.

Modules 2 and 3 are uniform since the tasks required of testees are of a considerably individual nature and not liable to cheating or copying.

The battery is accompanied by an administration and scoring

manual comprising ten sections to introduce the test to teachers and administrators and to guide the administration procedures, clearly and in detail. It also includes assessment scales, expected and/or acceptable answers to each task, the tape script, and data collection and feedback getting forms.

It could be said that the focus in this battery is on higher order language abilities. One of the fundamental requirements is to establish what the subjects can do with language at the specified level and how well they can perform defined tasks, rather than comparing them with one another. With reference to the syllabus design and the nature of the course, the type of tests which were judged to be more suitable were those provided by this battery i.e. direct types. However, for experimental reasons, a form of less direct type of test i.e. the Cloze was built into it for comparison and correlational studies.

Administration

The sample/the test population:

The battery was administered in April-May, 1983 to 460 male and female students (14-16 years) of 12 secondary schools in both Kuwait and the U.K. distributed as follows:

Number of Schools:	12	(10 in Kuwait and 2 in Britain)
Number of Subjects:	460	(370 in Kuwait and 90 in Britain)
Age group:	14-16 years	
Nationality groups:	3 Groups	(Kuwaitis, non-Kuwaiti Arabs and British)
Sex:	184 boys)	from Kuwait
	186 girls)	
	15 boys)	from Britain
	75 girls)	

The sample was randomly chosen in Britain and Kuwait, and test-taking took place on a purely voluntary basis on the part of the subjects. Students were told that it was an experimental test and would have no effect on their results at the end of the year. The pilot scheme generally aimed at testing the tests.

Administration of the battery in Britain actually aimed at 1. validating the test language, and 2. providing a model of performance and an inventory of the responses to the test tasks by native speakers of the same age group and presumably conceptual level as our subjects in Kuwait. Administration of the battery in Kuwait aimed at collecting adequate data for various validation and reliability studies. As the test had to be administered in each Kuwaiti school simultaneously, the preparation for this particular arrangement took a whole very busy week before the days assigned for actual test-taking by learners/testees. The actual administration of the battery was spread over three days, one for each of modules one and two and a day for both modules three and four.

Before actually administering the battery in Kuwait, it was introduced to all cadres of the Ministry of Education who are concerned with English language teaching. Meetings were held with the ELT Inspectorate to introduce the test, explain its rationale, discuss its format and components. They were invited to give their judgement of the battery as a whole, and of its relevance to the syllabus and course objectives and its suitability to the specified target level. Aware of the fact that any standardized written instructions can become no more than words on paper if procedures are unstandardized, meetings with Inspectors and later with head teachers and teachers in the ten schools of the pilot schemes were called. These meetings aimed at standardization of procedures to guard against possible indefinite individual interpretations of the administration and data collecting plan. Meetings were held as well with the officers of the Educational Technology Department of the Ministry of Education to check that the language laboratories in the pilot scheme schools were working, to provide tapes for recording students' performance and stand-by technicians as a precautionary measure against failure of any tape recorder in any laboratory. During the meetings, the battery constructor had to explain, more than anything else, why the test rubrics were introduced in both English and Arabic throughout the battery at that level. It was made clear that the aim is to neutralize any possibility of the learners misunderstanding the instructions. The point at issue is the performance of the task for which the testees are assessed.

After administration, more meetings were held for standardization of scoring procedures. The scoring system and assessment scale (see table 6 below) were discussed in detail, and the minutes of these meetings were taken for documentation of the event and the views of those present. On the basis of the views expressed in these trial scoring and standardization meetings, it was felt necessary to provide further detailed and more analytical scoring system.

TABLE 6. GENERAL ASSESSMENT SCALE FOR LEVEL SIX

ALTERNATIVE 1

Broad Grade	Range/	%	Result/ Decision	Detail
A	Excellent Achievement	91-100	Proceed to next level	- Distinctive achievement against the criterion performative level. - Message delivered appropriately and correctly even beyond the criterion level.

Broad Grade		Range	%	Result/ Decision	Detail
B	Good Achiever	B+	81-90	Proceed to next level.	- A very good achievement against the criterion performative level. Message delivered fairly well (appropriately and free from mistakes) just conforming within the criterion level, touching lightly the toleration line.
		B	71-80	Proceed to next level after short course	- A good achievement against the criterion performative level. Message conveyed reasonably clearly but with slight mistakes not much interfering with message comprehensibility. Something better is conceivable.
C	Satisfactory Fair	C+	61-70	Proceed to next level after detailed remedial course	- Satisfactory achievement against the criterion performative level. Message conveyed with some difficulty, hence understood with some difficulty, but partly distorted by some mistakes. Just about average.
	Modest	C	51-60	Borderline Case	- Needs re-consideration
D	Unsatisfactory achiever	D	41-50	Resit the test before long/next session	- Underachiever with reference to criterion performative level. Message obscured by mistakes which make it incomprehensible. Generally below satisfactory. Cannot be promoted to next level. Should resit the exam. the next session. Remediation needs to be intensive.
	Non-achiever/ Grossly underachieving	D-	-41	Resit the test after one year	- Non achievement or gross under achievement. No prospect of immediate remediation exists. Should re-do the course and consequently re-sit the test after one year.

Data Analysis plans and its aims

Analysis of data is underway to find out whether the battery tested what it was intended to test, how well the different subtests and/or tasks have performed and to examine the various aspects of test reliability as applicable to this battery. Inter- and intra-modules' correlational studies are planned as well as inter- and intra-rater reliability research. Statistical analyses and otherwise will also be carried out to determine internal consistency, particularly with the alternative form or the parallel test as applied to M1 & 4. Point bi-serial correlations will be carried out between performances of males and females, Kuwaiti and non-Kuwaiti Arabs, etc.

The battery as it stands is not a panacea. It is a serious attempt at contributing to the development of English language testing. Relevant analysis may determine its value and prove or disprove its validity and reliability. Consequently it will help to decide whether the approach is useful or otherwise; whether the attempt is worthwile or not; and which tasks to retain and which to discard from the final format.

It may be worthwhile to point out here that some aspects of validity are built from the outset into the test and throughout the construction and administration. As far as content validity is concerned the objectives of the course are the test criteria. The tasks were sampled from the syllabus; the course writers examined the battery and approved of it as testing their course; and the practising teachers and inspectors examined it and the majority admitted the considerable degree of coverage of the sample. With reference to face validity the battery was highly acceptable to its consumers. However, some aspects of validity are not applicable to this battery, e.g. concurrent and predictive validity because 1. no acknowledged standardized test of the same course exists so far for concurrent validation, and 2. it is not a proficiency test with a predictive outlook but an achievement test with retrospective perspective.

Conclusion

Compared with the current practices in test construction in Kuwait, the present battery breaks new ground by the very nature of its construction and arrangement of tasks in modules, themes and parallel sub-tests. It is hoped that post-pilot data analysis will prove its worth.

Notes

1 J.B. Carroll referred to this particular situation in the Arab Gulf States on various occasions. See in particular Carroll, J.B., 1982 "Is there another way?" In J.B. Heaton (ed)

"Language Testing", Modern English Publications Ltd., U.K., 1982.

2 The Committee Members were:-

- Dr R.R.K. Hartmann, Director of The Language Centre
- Mr G.C. James, the former Supervisor of the project
- Mr C. Weir, the current Supervisor of the project
- Mr T. O'Neill, one of the C.E.C. writers
- The Present Researcher

Bibliography

Baghdady, M.I. 1982a. "Modern Foreign Language Testing and the Current TEFL Situation in Kuwait". Unpublished paper presented at the Post-Graduate Seminar, The Language Centre, University of Exeter, March, 1982.

Baghdady, M.I. 1982b. "Current Issues in Testing". A paper presented at the London Workshop held in the Institute of Education London University, July 1982 sponsored by ELTA/O.U.P.

Baghdady, M.I. 1982c. "Prolegomena to Communicative activity-based Language Tests (CALT)". A paper presented at the ELTA/O.U.P. London Workshop held at the Institute of Education London University, July 1982. To appear in the ELTA/O.U.P. London Workshop publication.

Carroll, J.B. 1982. "Is there another way?" In J.B. Heaton (ed) "Language Testing", Modern English Publications Ltd., U.K., 1982.

Cronbach, L.J. 1964. "Essentials of Psychological Testing", Second Edition. Harper & Raw, N.Y., April, 1964: 30 and 108.

Morrow, K.E. 1979. "Communicative Language Testing: revolution or evolution?" In Brumfit & Johnson (eds) "The Communicative Approach to Language Teaching", O.U.P., 1979: 149-150.

O'Neill, T. & P. Snow. 1977. "Crescent English Course: Teacher's Book 1", p.1., ELTA/O.U.P., 1971, 1978, 1982.

THE FILL-IN TEST: A MODIFIED MULTIPLE-CHOICE CLOZE TECHNIQUE TO TEST READING COMPREHENSION OF ENGLISH AS A FOREIGN LANGUAGE

Marsha Bensoussan, Rachel Ramraz (University of Haifa)

1 Introduction

The Fill-in Test was developed by the writers over a period of eight years in an attempt to improve the efficiency of the English section of the Haifa University Psychometric Entrance Examination. The purpose of this paper is to describe the Fill-in Test, to explain how to construct it, to give some statistical evidence for its effectiveness, and to compare it with more conventional Multiple-Choice and Cloze Tests.

1.1 Disadvantages of conventional multiple-choice tests

One problem with constructing multiple-choice content questions is the lack of efficiency in time and materials: a student must read a great many lines of text in order to answer relatively few questions. A paragraph of approximately 100 words may yield, on the average, no more than five questions. Since a test's reliability depends, to a great extent, on the number of questions it contains, the results of a test of 20 minutes, which contains a page of text (approximately 400 words) and about 10 questions, are most often unsatisfactory. Too much time is spent on the input (reading and answering question), and not enough output (information about the student's reading ability) is obtained. In short, this method is somewhat inefficient. (1)

Another problem facing the test constructor is that as a result of pre-testing, the number of questions is usually reduced. A text must be accompanied by a minimum number of questions if it is to be useful. If, as sometimes happens after the elimination of unsuitable questions, too few good questions remain, the entire text cannot be used. If the text is to be salvaged, completely new questions have to be written and pre-tested, in the hope that the final number of questions will not fall below the minimum number.

Finally, it can be argued that the questions themselves constitute an extension of the text which is added to the student's burden during the test. The test mark, then, is a composite of understanding both text and questions and does not necessarily reflect comprehension of the text alone.

1.2 One solution: The Cloze procedure

1.2.1 Random Cloze

The appearance of the Cloze procedure offered a non-multiple choice

solution to the problems of efficiency and testing only comprehension of the text. The Cloze has been used for testing EFL reading comprehension as well as native comprehension (Oller 1973). Using Taylor's (1953) random Cloze procedure (i.e., testing a "random" sample of text), where the student is expected to write in his own words to fill the gap in the text, the test constructor can obtain one test item for every n words (usually every 7th word). The resulting Cloze test is likely to have a higher number of questions per line of text than the Multiple-Choice Test.

Stansfield and Hansen (1983) quote high correlations between Cloze scores and scores on other types of second language tests. Heilenman (1983) finds high correlations between random Cloze tests and other foreign language testing techniques. She advocates "the cautious use of a Cloze test as an alternative or supplement to other methods of placement" (p. 125).

Experimentation at Haifa University, however, showed that the random Cloze (i.e., the placement of a blank word after every nth word) did not suit our needs, and the procedure was rejected for the following reasons:

1. The spacing of blanks after every 7th word was too close, and students complained that they easily lost the thread of thought.

2. Some important words would not be tested if the mechanical choice of blanks was to be followed. Either the blank space would have had to be moved or the text rewritten.

3. The random Cloze cannot be pre-tested or altered. It is assumed that the information in each item is independent of that needed to answer other items (see Klein-Braley, 1981). That is, a response to a particular blank is determined by the context surrounding that blank, and not by other blanks in the text, except only indirectly. For this reason, if necessary and without harming the test as a whole, it should be possible to reduce the total number of test items by removing any number of blank spaces and replacing them by the original word of the text. Theoretically, if pre-test results indicate that a particular blank is found to be an unsuitable test item, then it can be replaced by the original word in the text, and the student can continue reading on to the next blank.

In the case of the random Cloze Test, however, it would not be possible to remove the unsuitable blanks at will. The unsuitable test item (i.e., blank space) would need to be re-filled by the original word in the text. The gap refilled, the number of words until the next blank space would then increase, for example, from 7 to 15.

A strict adherence to the rules of random Cloze would not permit the omission of any suitable test items. On the other hand, a more flexible attitude toward Cloze test construction would focus on the number, placement, meaning, and function of blank spaces in the entire passage rather than merely on the number of words between each blank space.

1.2.2 Non-random/Rational Cloze

Cranney (1973) and Alderson (1979) suggest that rational Cloze deletion would be more productive than the random Cloze. Although rational deletion would increase test construction time and result in a loss of objectivity, Greene (1965) also recommends its use.

A non-random Cloze procedure in which the deleted items mark relationships between propositions was developed by Levenston et al. (1982). They call this type of Cloze test "Discourse Cloze" because its deletions consist solely of cohesion markers on the macro-text level.

Assuming that the ultimate purpose of reading comprehension is a correct interpretation of marco-relationships, they argue that testing proficiency in the higher-level skill (macro-level) simultaneously tests proficiency at the lower (micro-) level as well.

They have adopted the distinction made by Van Dijk (1977, 1980) between micro- and macro-text levels. The former describes individual sentences or propositions in a text. Derived from the micro-structures of a text are the macro-structures, which represent intersentential, global or whole-text meanings of a discourse. According to Levenston et al. (1982):

> The inferencing process needed for reconstructing cloze tests can rely on a range of linguistic, pragmatic and textual clues.

For each of these categories, the following definitions are provided:

> By 'linguistic' we mean cases in which the meaning and form of the missing word are clearly specified by the linguistic micro-context in which the deletion occurs.

This 'linguistic' category is further subdivided into 'syntactic' and 'semotactic':

> By 'syntactic' we mean cases in which the word intended is clearly specified by the grammatical construction in which the deletion occurs. By 'semotactic' we mean cases in which the meaning and lexical choice of the intended word is confirmed (or, in the case of the cloze, provided) by the interaction of that word with the meaning and occurrence of other words within the context.

The 'linguistic' context does not exceed the sentence and therefore functions on the level of micro-processing.

In addition to the linguistic level, Levenston et al. (1982) discuss the pragmatic level of comprehension:

> We have termed the ability to replace items that require extra textual knowledge in the cloze a 'pragmatic' ability. At the most general level, pragmatic ability is the general knowledge of the world that provides the basis for forming expectations and interpreting given texts.

On the macro-level are the textual clues of cohesion:

> A third component of reading ... is the ability to follow the cohesive ties in a text in order to understand the interrelationship between the sentences. This ability differs from linguistic and pragmatic knowledge in that it is specific to text-processing as distinct from sentence-processing.

This textual component, which emdobies comprehension at the macro-level, includes the concepts of grammatical and lexical cohesion as described in Halliday and Hasan (1976). Grammatical cohesion includes the features of anaphoric reference, substitution, ellipsis, and conjunction; lexical cohesion includes the features of repetition, hyponyms (and super-ordinates) antonyms, synonyms, and collocations (on the macro-level).

In contrast to the "Discourse Cloze" proposed by Levenston et al. (1982) which places blanks only instead of those words tapping high-level (macro-) skills, the Fill-in Test includes the entire range of possible clues on all three levels: linguistic, pragmatic and textual. We nevertheless adopted the theoretical framework described by Levenston et al. (1982) in explaining the construction of the Fill-in Test.

We had previously used the modified rational open-ended (or free response) Cloze for the English Psychometric Examination at Haifa University with good statistical results. (See Bensoussan 1981 and below, EXPERIMENT THREE). However, because the marking was inefficient, it did not solve our problem completely.(2) We were forced to return, rather, to the multiple-choice format for purely technical reasons.

1.3 Other solutions: Multiple-Choice modification of the random Cloze procedure

The idea of combining both Cloze and Multiple-Choice procedures in one test is not new. Ozete (1977) discusses Ronald Carver's "reading in-put test" which has two alternate responses. Guthrie (1973) introduced "the maze task" -- a Cloze passage with three alternate responses for each blank. Cranney (1973), Porter (1975), and Jonz (1976) discuss Cloze tests with four possibilities for every blank. According to Cohen (1980: 94-95), the addition of the multiple-choice technique to the Cloze is a major change; the student performs a different task than in the open-end Cloze.

In part, some of the problems of multiple-choice techniques are re-

introduced. The student must choose the best alternate response on the basis of his comprehension of the text. It is the distractors and not necessarily the text that can make the test difficult (Bensoussan, 1982; Pikulski and Pikulski, 1977). In spite of this disadvantage, however, the multiple-choice Cloze is recommended because it is short, and at the same time, it has a high reliability (Jonz, 1976).

1.4 Our solution: the Fill-in Test

We based the Fill-in Test on a modified version of the Cloze: words were deleted on the basis of a rational, not a random process. The rational Cloze was preferred because it gives the test constructor more control over specific areas of comprehension and over the items to be tested and diagnosed. The test can be tailor-made to suit a particular group of students or teaching points.

Since our basis was not strict, random Cloze, and since our test closely resembled a Multiple-Choice Test in format, we decided to call it a Fill-in Test, so as not to confuse it with standard Cloze.

We used the procedure of inserting 20-30 blank spaces into a 300-word text. Each blank space takes the place of a word or phrase (group of not more than three words), and for each blank space, there is a choice of 4 possible answers. The Fill-in Test modifies the Cloze procedure in three ways:

1. Possible responses are already provided in a multiple-choice format -- unlike the Cloze, which contains gaps in the text that need to be filled in.

2. Unlike the Cloze test, a blank space in a Fill-in Test can take the place of more than one word.

3. Blank spaces are placed not after every nth word, but within a range of 7 to 15 words or more.(3)

2. The Fill-In Test

2.1 Linguistic basis

It was assumed that a given blank space could tap one of two levels of reading: at the micro-level -- specific understanding of a word or collocation where the clue(s) appeared in close proximity to the blank (one or two words before and/or after), and at the macro-level -- a more general understanding of the whole test (e.g., writer's opinion, key words showing comprehension of concept and contrast/opposition , main idea of paragraph).

If we adopt the theoretical framework described by Levenston et al. (1982), we can analyze Fill-in Test A (see Appendix) as a sample test to

show the distribution of the test items according to both micro-, pragmatic, and macro-levels of reading. Since this was a real test and not just a theoretical exemplification of an ideal Fill-in Test, many of the features in the theoretical framework did not actually appear as part of the test and will therefore not be discussed in this paper. All actual items on the test, however, will be related to the theoretical model.

TEXTUAL CLUES	test item(s)
1 linguistic (micro-)	
1.1 syntactic	L
1.2 semotactic	
1.2.1 words or phrases (distractors chosen from words already in the text)	
1.2.1.1 nouns	E,H,M
1.2.1.2 adjectives	S
1.2.1.3 verbs	O
1.2.1.4 phrases	N, F
1.2.2 collocations	
1.3 both syntactic and semotactic	
1.3.1 parts of speech (same root, give different forms), test syntax and understanding of the relationship between words	K
1.3.2 word forms, as they affect comprehension	
1.3.2.1 adjectives	C,D
1.3.2.2 verbs	J
1.3.2.3 pronouns	A,I
1.3.2.4 prepositions	R
1.3.3 collocations	C,R
2 pragmatic (micro-)	H,T,U
3 textual (macro-) -- cohesion	
3.1 grammatical	
3.1.1 anaphoric reference	A,I
3.1.2 conjunctions	G,P
3.2 lexical	
3.2.1 superordiates	D,E,M
3.2.2 collocations	D

Some test items touch on more than one of the text components, either because of the multiplicity of clues in the text or because of the possibilities presented by the alternate responses (distractors).

2.2 Fill-in Test construction

2.2.1 Basic assumptions in choosing blanks

The basic criterion for choosing blank spaces is that there is enough redundancy in the text so that a proficient reader could use the clues to fill in the gap with an appropriate word or expression. Moreover, in

choosing the blanks, the test constructor should focus on pivotal or key words in a logical argument to see whether a student can follow the thought sequence.

Since it is relatively difficult to find words that indicate a student's knowledge of the whole text, preference could be given to first placing blanks which test reading comprehension. Function words, such as "however" and "therefore", would be good places for blanks. Content words such as nouns, adjectives, and verbs which carry the weight of an argument would also be useful, and their opposites would be included among the distractors. In this way, the test constructor may suggest alternate misleading logical thought sequences, but only one set of choices would be consistent with the writer's intentions within the text as a whole.

Other items tested could be cohesive markers such as "not only ... but also," "either ... or," and "on one hand ... on the other hand." It is assumed that a student's recognition of these syntactic devices would enable him to follow the flow of an argument, and that lack of recognition would impede his comprehension.

We aimed for a wide variety or sampling of test items. For this reason half the blanks represent content words (nouns, verbs, adjectives, adverbs) and the other half function words (conjunctions, prepositions, word forms). For examples, see Fill-in Test A, Appendix. Test constructors are free, however, to focus the test on whatever level of discourse they wish to test.

When pre-testing, it is advisable to place approximately 50% more blank spaces than needed even though shrinkage is usually minimal. For example, if 15 items are required for a test, 20 or 30 may be pre-tested. Afterwards, when unsuitable test items are eliminated, these gaps can either be filled with the original word, or else the whole phrase in which it appears (provided it is not a key phrase) may be eliminated. The Fill-in Test still remains intact, because each item is independent of the others (see Alderson, 1979; Jonz, 1976; Klein-Braley, 1981), even after many of the items have been eliminated.

2.2.2 Rationale for determining alternate responses (distractors)

In thinking up alternate responses (distractors), the test constructor would be expected to use words focusing on a particular point, either in terms of content or structure. It was found, in fact, that items focusing on a single point (e.g., in which students were to choose from among 4 adjectives) were more successful than items unclearly focused (e.g., alternate responses included 2 adjectives and 2 conjunctions, so the student does not understand if he is to look for content or form or both; another even worse possibility: 1 adjective, 1 noun, 1 conjunction, and 1 relative pronoun). The student probably does not understand the point that is being tested, and therefore the results would not really show whether the student has understood the text. (Examples of less

desirable items in Fill-in Test A would be B and F.)

In choosing distractors, the test constructor may make use of collocations, presenting words that could appear together and make sense in some other context. For this reason, opposites are particularly useful: they test the student's understanding of the whole text. Conjunctions are also helpful here. A student choosing "therefore" when only "however" would fit the context may have understood a particular sentence, but certainly did not grasp how the sentence fit into the context as a whole.

For testing English as a foreign language, synonymous distractors should not be used. It is advisable to avoid distractors where the correct choice is ambiguous even for native speakers. Thus, one should also avoid asking about detailed grammatical points (e.g., the distinction between it's and its) or prepositions which may also be confused by native speakers. The Fill-in Test is essentially a test of reading comprehension, not of grammar.

If placement of the blank spaces is based on a linguistic examination of the text rather than randomly, it might be argued that one way of constructing the Fill-in Test would be to pre-test it as a Cloze. Likely blank spaces could be chosen by the test constructor, and the possible distractors, it might be supposed, might be found from among the students' wrong answers (see Jonz, 1976).

Experience shows, however, that only about one-third of the test items can be obtained in this way. This method is very time consuming and yields relatively little in return.(4) The best way to construct the Fill-in Test is to decide before-hand the structures and ideas that are to be tested and to place blanks where these points are likely to be tapped.

2.3 Comparison between the Fill-in Test and the Multiple-Choice Test

Having developed the basic outlines of the Fill-in Test, we wanted to verify that the Fill-in Test would do its job as well as the more well-known, conventional Multiple-Choice Test. Four experiments were carried out, using four separate test batteries which compared test items and scores of Fill-in Tests with those of Multiple-Choice Tests. Number of items per line of text, difficulty levels of items, extent and function of test questions, correlations between scores, and test reliability were examined.

3 Experiments

3.1 Experiment One: Length, Difficulty, and Reliabilitiy

3.1.1 Procedures

The sample consisted of 435 first-year students taking the advanced

reading course in English as a Foreign Language at Haifa University in 1973. Each student took one of three different English tests consisting of 4 subtests: 1 Fill-in Test and 3 Multiple-Choice (M-C) Tests (texts accompanied by multiple-choice content questions as well as by vocabulary and reference questions).

3.1.2 Results

For a description of EXPERIMENT ONE, see Table 1 describing 3 English tests.

Table 1

Description of 3 English Tests

Test**(N=435)	English Test 1 ($n=154$)			English Test 2 ($n=170$)			English Test 3 ($n=111$)		
Subtest	number lines	number items	$\overline{ET}$*	number lines	number items	$\overline{ET}$*	number lines	number items	$\overline{ET}$*
(1) Fill-in	30	28	52%	27	28	56%	32	24	55%
(2) M-C A***	5	6	48%	7	7	50%	12	6	63%
(3) M-C B	18	11	54%	13 6	6	42%	14	7	43%
(4) M-C C	24	8	51%	26	11	43%	21	10	50%
Total	77	53		73	52		79	47	
Reliability:									
Kuder-Richardson			.828			.709			.761
Split-Half			.8669			.6113			.7963

*$\overline{ET}$ = average of the Easiness Indices (percent of correct responses out of all the attempted answers per item)

** N = number of subjects

*** The relatively large number of items on this test is a result of asking mostly word meaning and not content questions.

In terms of easiness, Table 1 indicates that Fill-in items are on a par with those of the conventional multiple-choice tests. Although the Fill-in Test constituted little more than one-third the total number of lines of text, it contributed at least one-half of the total number of test items.

The reliability of each of the tests was fairly high.

3.2 Experiment Two: Text Clues, Difficulty, Discrimination, Reliability

A question arises concerning the difference between the two types of test. Although they may yield similar statistical results, one could not go as far as to say that the Fill-in and Multiple-Choice (M-C)

Tests examine the same skills. Nevertheless, both test the reading comprehension of a particular text. In order to have a better basis for comparison, it was decided to take 4 texts and test each text twice using a different format and different students each time. Each test was constructed first in the conventional multiple-choice format, and the second time, the same text was used to construct the new Fill-in format.

3.2.1 Procedures

A total of 1487 applicants to the first-year of studies at Haifa University were tested. Most are high school graduates who have had seven to eight years of English. At random, each student received one text with questions.

This test battery consisted of eight tests; that is, 4 separate texts -- once as an M-C Test and once as a Fill-in Test. The texts ranged in length from 28 to 36 lines; there were 9 - 15 questions on the M-C Tests and 20 - 28 questions on the Fill-in tests. (For a sample Fill-in and Multiple-Choice text pair, see Fill-in and M-C Tests A, Appendix).

3.2.2 Textual clues

In order to examine the extent to which each of the two formats focussed on macro- and micro-levels of comprehension, we analysed each test item of Test A (both M-C and Fill-in test versions).

Breaking down the results according to our criteria for finding textual cluse (see page), i.e., micro-level (linguistic and pragmatic) and micro-level (textual) clues), we see that the distribution of textual clues are similar for both M-C and Fill-in Tests. As would be expected, however, the Fill-in Test has more micro-level linguistic test items. (See Table 2, which breaks down the items according to the amount of text included.)

Table 2

Breakdown of Items According to Amount of Text Included:
M-C and Fill-In Tests A

Category	Number of Items:	Fill-in Test	M-C Test
1.	Linguistic (micro-level)	12	7
2.	Pragmatic (micro-level)	2	1
3.	Textual (macro-level)	7	7
	Total No. of Items	21	15

The Fill-in Test need not be restricted to comprehension of single words; it can be used to tap an understanding of the wider context, in proportions similar to those of the M-C Test. In both, the most frequent kind of item included micro-linguistic and macro-textual clues, although the Fill-in Test had more linguistic clued and the M-C Test had proportionately more textual clues.

3.2.3 Results

A comparison between the Fill-in and M-C appears in Table 3 and shows that the means of the raw scores are almost equivalent. The Fill-in Test yields more items, and its reliability is higher. The M-C Tests, however, have a higher percentage of good questions(i.e., Discrimination Index greater than .30) than the Fill-in Test.

Since there are greater differences among the average raw scores for Texts A, B, C, and D than there are between the M-C and Fill-in versions of each text, the results would indicate that the choice of text may be more important than the format by which it is tested.(5)

Table 3

A Comparison Between Fill-In and Multiple-Choice Tests on Same Text

	Text A		Text B		Text C		Text D	
	Fill-in	M-C	Fill-in	M-C	Fill-in	M-C	Fill-in	M-C
No. lines	30	30	28	28	30	30	36	36
No. students	73	192	204	180	201	213	213	211
No. items	21	15	28	11	24	9	20	13
No. good items Disc. Index > .30(%)	18 (86%)	15 (100%)	23 (82%)	10 (91%)	22 (92%)	9 (100%)	14 (70%)	12 (92%)
*Median Disc. Index	.46	.50	.41	.49	.44	.45	.42	.42
Reliability:								
Kuder-Richardson(20)	.80	.76	.79	.63	.80	.58	.69	.62
Split-Half	.80	.78	.78	.62	.80	.48	.67	.58
score (%) and standard deviation	63% 18.35	62% 16.56	6 % 14.45	64% 18.41	75% 17.26	72% 11.43	46% 19.50	44% 14.86

* Discrimination Index = point-biserial correlation between the response and the student total raw test scores. An item is considered effectively able to discriminate between good and weak students if its Discrimination Index is greater than .30 but less than .60.

3.3 Experiment Three: Two test batteries containing both Fill-in and Multiple-Choice items: Difficulty, Reliability, Pearson Correlations

3.3.1 Procedures

The Fill-in Test was compared with the Multiple-Choice Test in two additional test batteries, each consisting of Fill-in and Multiple-Choice test items according to the following format:

Multiple-Choice Subtest A - very easy text
Multiple Choice Subtest B - fairly easy text
Multiple Choice Subtest C - text of medium difficulty
Fill-in Subtest D - difficult text
Multiple-Choice Subtest E - very difficult text

The texts were graded from A, the easiest, to E, the most difficult. For convenience, it was decided to consider all four Multiple-Choice texts as a combined single long text when comparing results with those of the Fill-in. The first test battery contained 56 items and the second 55.

The test batteries were administered as the English section of the Entrance Examination to the Universities of Haifa and Tel Aviv during two consecutive years. The first year, the examination was administered to 7,499 applicatns; the second year, it was administered to 7,114 applicants.

3.3.2 Comparison between Fill-in and Multiple-Choice Tests -- Results

The results of the two test batteries, which appear in Table 4 below, show the comparison between Fill-in items and Multiple-Choice questions in each test battery.

Table 4

A Comparison Between Fill-In and Multiple-Choice Tests

Year	Number of Subjects	Total Number of Test Items	Number of Items	Mean Score (%)	Stand. Dev.	K-R Rel.*	Pearson Correlations Between Total Scores of Fill-in and Multiple-Choice Tests
1980	7499	56	Fill-In 15	9.0(60%)	3.98	.841	.748
			M-C 41	21.5(52%)	8.75	.903	
1981	7114	55	Fill-In 17	8.5(50%)	4.32	.815	.798
			M-C 38	22.0(58%)	8.39	.880	

*Average Kuder-Richardson #20 reliability for two parallel test forms.

Mean scores of the Fill-in Test were similar to those of the Multiple-Choice questions. Apparently, Fill-in items can be made easy or difficult by the test constructor, in the same way that Multiple-Choice questions can be graded for difficulty.

As for Pearson correlations between total Fill-in and Multiple-Choice scores, for the first test battery it was .75, and for the second it was .79. These figures are considerably higher than those obtained in PART FOUR of this study (see below), where the Pearson Correlation ranged from .36 to .47. It must be remembered, however, that the correlations obtained in PART FOUR are based on only ten multiple-choice questions, whereas the present correlations were based on 38 and 41 multiple-choice questions, respectively.

When correlations are not exceedingly high, the reason is usually that there is an external factor common to both tests. Each test, then, gives information of a different kind about the student's reading comprehension. In this test battery, the Fill-in subtests did not correlate so highly as to permit their substitution for the Multiple-Choice subtest. Since each test contributes another measure of information about comprehension, it would be most efficient to use <u>both</u> Fill-in and M-C Tests in the complete test battery.

3.4 Experiment Four: Difficulty, Pearson Correlations

In assessing the Fill-in, we wished to compare it with other types of reading comprehension tests. Accordingly, the three Fill-in subtests described in Experiment One (above, p.) were compared with another test battery which had been previously administered to 354 applicants to Haifa University. (See Bensoussan, 1981: Kuder-Richardson No. 20, Reliability = .93.) The same students took both the Multiple-Choice/Cloze Test battery and one of the three Fill-in subtests.

3.4.1 Test Battery

The English section of the Entrance Examination consisted of 57 test items and was 75 minutes in duration. Each of the items was selected for level of difficulty and discrimination by pre-testing a similar population at Haifa University.

Within the multiple-choice framework, many types of testing exercises are possible. We used the following three multiple-choice subtests:

3.4.1.1 <u>Sentence Completion subtest</u>, which was a test of word form (i.e., content words) and syntax (i.e., function words) where the student chose the word(s) that best completed the sentence.

3.4.1.2 <u>Vocabulary Substitution subtest</u>, a test in which the student was asked to find the best synonym for the underlined word in each sentence.

3.4.1.3 <u>Multiple-Choice Comprehension subtest</u>, a text accompanied by

multiple-choice questions about content, syntax (conjunction), vocabulary, and reference. (Same type as above, Experiments One, Two and Three).

3.4.1.4 The Cloze subtest consisted of a 313-word text containing 26 blank spaces. Each student response was marked either correct or incorrect according to comprehension (i.e., whether it was clear that the student understood the meaning of the context). Spelling errors were not counted. A panel of 12 teachers graded the examinations, and acceptable responses had to be agreed upon unanimously during the marking of papers. It was assumed that someone who was able to fill in the gaps in the text demonstrated the ability to read and understand the passage.

Sentence Completion and Vocabulary Substitution, the first two subtests, were short, consisting of only one or two sentence, whereas the Multiple-Choice Comprehension and Cloze subtests presented much longer and more complex reading passages.

3.4.2 Results

3.4.2.1 Hierarchy of difficulty

The Cloze and the Fill-in Tests together have the greatest number of questions (26 Cloze + 28 or 24 Fill-in). If we compare the Fill-in with each of the subtests in the test battery in terms of their respective difficulty levels (see Table 5), we obtain the following hierarchy: Vocabulary is the easiest subtest. It is followed by the more difficult Fill-in, Cloze, Multiple-Choice, and Sentence Completion subtests, all three of which are approximately of equal difficulty. This general pattern appears for each of the three Fill-in subtests.

Table 5

Comparison of Difficulty Levels of A M-C Test Battery and Three Fill-in Subtests

	FILL-IN SUBTEST			MULTIPLE-CHOICE BATTERY + CLOZE							
				Sent. Comp.		Vocabulary		M-C Text		Cloze	
N	Form	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
119	1	14.5	4.4	5.8	2.8	5.5	2.5	4.5	2.0	13.2	6.7
142	2	14.3	3.1	5.9	3.0	5.7	2.4	4.6	2.4	13.9	6.1
93	3	13.1	3.1	5.6	2.9	5.5	2.4	4.4	2.3	13.4	6.6
Total-across all 3 subtests:											
354		14.1	3.6	5.8	2.9	5.6	2.4	4.5	2.2	13.5	6.4
Mean %		54%		48%		62%		50%		52%	
No. items in each subtest Fill-in Form:											
Subtest:		1) 28		12		9		10		26	
		2) 28									
		3) 24									

3.4.2.2 Correlations between subtests

In general, the subtests do not correlate with each other (See Table 6), although all correlations obtained were significant. The highest correlation (.64) was bound between scores of the Sentence Completion Subtests and the Multiple-Choice Test. The relatively low correlations would indicate a relation with some larger common factor, such as the testing of EFL reading comprehension. The lowest set of correlations was between Vocabulary and M-C. Correlations between these subtests were not so high as to permit the possibility of substituting one for another. In this respect, each subtest appears to be tapping a different area of reading comprehension.

It is especially interesting to note that the Fill-in subtests, multiple-choice versions of the modified Cloze procedure, do not correlate highly with either the Multiple-Choice Test or the rational Cloze passage. Thus one may conclude that the Fill-in Test might be testing something different.

Table 6

Pearson Correlations Between Subtest Scores

Subtest	Fill-in	Sentence Completion	Vocabulary	Multiple-Choice	Cloze
Fill-in					
***TOTAL	1.00	.50	.40	.41	.43
Subtest 1		.50	.38	.36	.38
Subtest 2		.53	.38	.47	.50
Subtest 3		.50	.46	.45	.48
Sentence Completion					
TOTAL		1.00	.43	.64	.54
Subtest 1			.39	.62	.54
Subtest 2			.50	.66	.60
Subtest 3			.35*	.64	.48
Vocabulary					
TOTAL			1.00	.34	.40
Subtest 1				.27**	.41
Subtest 2				.38	.42
Subtest 3				.35*	.36
Multiple-Choice					
TOTAL				1.00	.53
Subtest 1					.52
Subtest 2					.64
Subtest 3					.39
Cloze					
TOTAL					1.00
Subtest 1					
Subtest 2					
Subtest 3					

All correlations are significant, and all are $p = .0001$ except * $p < .0001$ and ** $p < .001$

***TOTAL: across all three Fill-in subtests.

4. In Conclusion

The purpose of the Fill-in Test is not to replace the conventional M-C Test but to offer an additional multiple-choice test format. The Fill-in tests reading comprehension -- not only words and word forms at the micro-level, but, more importantly, the ability to follow a logical thought sequence at the macro-level of reading.

Statistically, the Fill-in Test measures up to the traditional M-C Test. A test constructed with the Fill-in format will probably have items of the same average difficulty and effectiveness as if it had been constructed in the conventional M-C format. The only differences are that it will probably have more test items (administered in the same amount of time), and therefore the reliability will be slightly higher. The Discrimination Indices, however, will probably be lower.

The Fill-in Test has some other advantages: it is easier to construct than the M-C Test. Each item tests only one point and contains only one word or short phrase. Moreover, after pre-testing has eliminated unsuitable test items, the Fill-in Test is more likely to remain intact than the M-C Test, for which it is sometimes difficult to find enough test items.

For these reasons, we use a Fill-in subtest in our English examination. By permitting the inclusion of a greater number of test items in the same amount of time as one M-C test, the Fill-in Test helps increase the reliability and improve the quality of the test battery as a whole.

Notes

1 The other subtests in Haifa University Psychometric Examination yield 25-40 test items for each 30-minute test session.

2 It was inconvenient to mark thousands of examinations by hand and to constantly re-assess standards of acceptability. Approximately 2,000 papers were marked by 20 teachersin a large room during an 8-hour marking session. Every questionable answer was put to the whole group of markers, so that marking could be consistent for every test marker. This discussion, as well as corrections and changes that resulted from the discussion, added to the marking time.

3 Blanks do not represent a random sampling of the text; each blank focuses on a specific semantic or syntactic point, making the Fill-in a discrete point test.

4 It was suggested by Dr. Valerie Whiteson, English Department, Bar-Ilan University, that a pre-test using the Cloze would be successful provided that the English proficiency of the students was high enough, that is, near native level (personal communication).

5 The amount of pragmatic knowledge required for comprehension was a vital factor affecting differences in students' reading, according to Levenston (personal communication).

Bibliography

Alderson, J. 1979. "The Cloze procedure and proficiency in English as a foreign language". TESOL Quarterly, 13, No. 2: 219-227.

Bensoussan, M. 1981. "A comparison of Cloze and Multiple-Choice reading comprehension tests". Report No. 57. Haifa: University Selection and Assessment Unit.

- 1982. "Testing the test of advanced EFL reading comprehension: To what extent does the difficulty of a multiple-choice comprehension test reflect the difficulty of the text?" System, 10, No. 3: 285-290.

Cohen, A. 1980. Testing Language Ability in the Classroom. Rowley, Mass.: Newbury House.

Cranney, A.G. 1972-73. "The construction of two types of Cloze reading tests for college students". Journal of Reading Behavior, 5, No. 1: 60-64.

Greene, F.P. 1965. "Modification of the Cloze procedure and changes in reading test performances". Journal of Educational Measurement, 2, No. 2: 213-217.

Guthrie, J.T. 1973. "Reading comprehension and syntactic responses in good and poor readers". Journal of Educational Psychology, 65, No. 3: 294-299.

Halliday, M.A.K. and Ruqaiya, H. 1976. Cohesion in English. Longman: London.

Heilenman, L. 1983. "The use of a Cloze procedure in foreign language placement". Modern Language Journal, 67, No. 2: 121-126.

Jonz, J. 1976. "Improving on the basic egg: The M-C Cloze". Language Learning, 26, No. 2: 255-265.

Klein-Braley, C. 1981. Empirical Investigations of Cloze Tests: An Examination of the Validity of Cloze Tests as Tests of General Language Proficiency in English for German University Students. Ph.D. dissertation, University of Duisburg, Germany: Liverpool.

Levenston, E.A., R. Nir, and S. Blum-Kulka. 1982. "Discourse analysis and the testing of reading comprehension by Cloze techniques". Paper presented at the International Symposium of Language for Special Purposes, Eindhoven, The Netherlands.

Nevo, B., E. Shor and R. Ramraz. 1975. "ITANA V, User's Guide: A Fortran program for multiple-choice tests and item analysis". Report No. 22, Haifa: University of Haifa Selection and Assessment Unit.

Oller, J. Jr. 1972. "Scoring Methods and difficulty levels for Cloze tests of proficiency in English as a second language". Modern Language Journal, 56, No. 3: 151-158.

- 1973. "Cloze Tests of second language proficiency and what they measure". Language Learning, 23, No. 1: 105-118.

- and C.A. Conrad. 1971. "The Cloze technique and ESL proficiency". Language Learning, 21, No. 2: 183-195.

Ozete, O. 1977. "The Cloze procedure: A modification". Foreign Language Annals, 10, No. 5: 565-568.

Pikulski, J.J. and E.C. Pikulski. 1977. "Cloze, maze, and teacher judgment". The Reading Teacher, 30, No. 7: 766-770.

Porter, D. 1975. "Modified Cloze procedure: A more valid reading comprehension test". English Language Teaching Journal, 30, No. 1: 151-155.

Ramraz, Rachel. 1979. "ITANA V: Computer program for item analysis". Report No. 40. Haifa: University Section and Assessment Unit.

Stanfield, C. and J. Hansen. 1983. "Field dependence-independence as a variable in second language Cloze test performance". TESOL Quarterly, 17, No. 1: 29-38.

Taylor, W.L. 1953. "Cloze procedure: A new tool for measuring readability". Journalism Quarterly, 30: 415-433.

- 1956. "Recent developments in the use of the Cloze procedure". Journalism Quarterly, 33: 42-48, 99.

Van Dijk, A. 1977. Text and Context. Longman: London and N.Y.

- 1980. Macrostructures. Lawrence Erlbaum: Hillsdale, N.Y.

Appendix

Fill-In Test A (Experiment Two)

Poor vs. Rich: A New Global Conflict

A conflict between two worlds -- one rich, one poor is developing, and the battlefield is the globe itself. On one side are two dozen or so industrialized, non-Communist states ____(A)____ 750 million citizens consume most of the world's resources, ____(B)____ most of its manufactured goods, and enjoy history's ____(C)____ standard of living. On the other side, demanding an ever larger share of that

____(D)____ , are about 100 underdeveloped poor ____(E)____ with 2 billion people -- millions of whom ____(F)____ in the shadow of death by starvation or disease. ___(G)___ , the conflict has been limited to economic pressure and proposals, and ____(H)____ in international forums. But the needs of the underprivileged nations are so pressing that some Western politicians describe ____(I)____ as a "time bomb for the human race." The conflict ____(J)____ the international economic system on which the ____(K)____ of much of the world is based.

In the U.N. General Assembly, ___(L)___ they are now a solid voting bloc, the developing states have approved resolutions that demand a "new international ___(M)___ order." The meaning: massive and painful sacrifices by the rich ___(N)___ the poor. So one-sided have the Assembly's actions become that the U.S. has ___(O)___ them as "a tyranny of the majority."

___(P)___, the U.S. along with other First World nations, admits ___(Q)___ there is a real grievance behind the angry rhetoric. Although the Third World population is literally exploding -- there are 200,000 new mouths ___(R)___ feed every day -- the land available for growing food is ___(S)___. As life in the countryside worsens, millions of peasants abandon their ___(T)___ and head for the slums of the developing world's cities, vainly ___(U)___ jobs that do not exist. Widespread poverty is a problem that afflicts all underdeveloped countries.

	1	2	3	4
A.	what	whose	whom	which
B.	produce	but	for	occupation
C.	low	lowest	high	highest
D.	poverty	poor	wealth	rich
E.	people	states	living	industry
F.	to	great	exist	worried
G.	Finally,	However,	After which,	So far,
H.	speeches	industry	states	producing
I.	him	them	it	those
J.	must destroy	destroyed	could destroy	to destroy
K.	stabilized	stabilizer	stable	stability
L.	and	where	that	however
M.	economic	cultural	produce	politics
N.	in addition to	as a result of	on behalf of	in exchange for
O.	said	denounced	praised	told
P.	Furthermore,	Nonetheless,	Therefore,	Since,
Q.	in	by	to	that
R.	in	for	to	and
S.	diminishing	increasing	more fertile	less dry
T.	money	wives	towns	farms
U	new	tried	seeking	finding

Appendix

Multiple-Choice Test A (Experiment Two)

A conflict between the worlds -- one rich, one poor is developing, and the battlefield is the globe itself. On one side are two dozen or so industrialized, non-Communist states whose 750 million citizens consume most of the world's resources, produce most of its manufactured goods, and enjoy history's highest standard of living. On the other side, demanding an ever larger share of that wealth, are about 100 underdeveloped poor states with 2 billion people -- millions of whom exist in the shadow of death by starvation or disease. So far, the conflict has been limited to economic pressures and proposals and speeches in international forums. But the needs of the underprivileged nations are so pressing that some Western politicians describe them as a "time bomb for the human race." The conflict could destroy the international economic system on which stability of much of the world is based.

In the U.N. General Assembly, where they now constitute a solid voting bloc, the developing states have approved resolutions that demand a "new international economic order." The meaning: massive and painful sacrifices by the rich on behalf of the poor. So one-sided have the Assembly's actions become that the U.S. has denounced them as "a tyranny of the majority."

Nonetheless, the U.S., along with other First World nations, admits that there is a real grievance behind the angry rhetoric. Although the Third World population is literally exploding -- there are 200,000 new mouths to feed every day -- the land available for growing food is diminishing. As life in the countryside worsens, millions of peasants abandon their farms and head for the slums of the developing world's cities, vainly seeking jobs that do not exist. Widespread poverty is a problem that afflicts all underdeveloped countries.

Textual Clues
1. micro-linguistic
2. micro-pragmatic
3. macro-textual (cohesion)

Questions on Multiple-Choice Text A: — Textual Clues

A. The aim of this passage is to
   1. explain a theory
   2. criticize a philosophy
   3. argue a cause
   4. present a problem — (3)

B. The industrialized, non-Communist states (line 3) are the world's
   1. richest
   2. poorest
   3. more populated
   4. least developed — (1)

C. or so (line 3). These words mean
   1. either
   2. about
   3. in that way
   4. therefore (1)

D. its (line 5). This word refers to
   1. citizens
   2. manufactured goods
   3. the world
   4. resources (3)

E. According to lines 9-11, so far the underprivileged nations
   1. have done very little to help themselves
   2. have been in constant armed conflict
   3. have no demands at all on the richer nations
   4. have already put some pressure on the richer nations (1)

F. Some western politicians (line 12) think that the under privileged nations
   1. will have to wait until the richer nations can help them.
   2. will continue making speeches and political agreements in order to fulfill their needs
   3. will use violence unless their needs are fulfilled
   4. will have to realize that they will always remain poor (1)

G. According to lines 10-15, the conflict is dangerous because
   1. it could upset the economic stability in the world
   2. the poor nations might begin a nuclear war
   3. the rich nations might begin a nuclear war
   4. all the nations would become poor in the end (1)

H. In the U.N. General Assembly, new resolutions demand that
   1. the rich nations give much more to the poor nations
   2. the rich nations approve of more poor nations
   3. the poor nations sacrifice more for the rich nations.
   4. the poor nations make more massive efforts for the rich nations (1)

I. According to lines 20-21, the U.S. thinks that the Assembly's resolutions
   1. are fair to all nations
   2. are in the interest only of the rich nations
   3. are in the interest of only the poor nations
   4. are insufficient in the fact of mass starvation (2)

J. The word Nonetheless (line 22) is followed by a sentence which
   1. agrees with the idea found in lines 20-21
   2. contrasts with the idea found in lines 20-21
   3. gives an example of the idea found in lines 20-21
   4. explains the idea found in lines 20-21 (3)

K. According to the author, the U.S. attitude towards the poor nations is that
   1. they have a right to ask for help, although their language may be too violent.
   2. they have a right to ask for help, and their language is not violent enough
   3. they have no right to ask for help, since they are only exaggerating the real situation
   4. they have no right to ask for more help, since they are already getting the maximum that the First World is able to give (3)

L. vainly (line 28). This word means
   1. uselessly
   2. proudly
   3. conceitedly
   4. irreverently (1)

M. According to lines 22-30, peasants of the Third World Population who leave their farms
   1. find a higher standard of living in the cities
   2. develop the world's cities
   3. diminish the land available for growing food
   4. find no better life in the cities (3)

N. The author of the passage implies that
   1. industrialization in a country probably leads to corruption
   2. a decrease in population growth would help solve the problem of poverty
   3. country life is much healthier than city life
   4. the rich nations don't really want to help the poor nations (3)

O. The main idea of this passage is that
   1. the rich and poor are in constant conflict
   2. the industrialized non-Communist states are rich and powerful
   3. poverty is a problem that must be solved by rich and poor nations alike
   4. peasants are migrating from country to city in search of a better life (3)

SELF-ASSESSMENT IN ENGLISH AS A FOREIGN LANGUAGE: SOME PRELIMINARY IDEAS

Nelson A. Berkoff (The Hebrew University, Jerusalem)

1 Some Preliminary Ideas

We spend a great deal of time preparing, marking and going through tests, but the one person who should be the most interested in the whole process - the student - is rarely asked to evaluate himself or even to have a say in the construction of tests. If we consider a normal teaching situation there are various points of time when a good, reliable, valid test is required - at the beginning and end of the course and at certain intervals along the way, when both students and teacher are interested in finding out what achievement or progress has taken place.

Consequently, I decided to investigate the possibility of constructing Self-Assessment Tests which might be suitable for all or any of these various points of time and, in particular, I had in mind the possibility of a reliable Self-Assessment Test that could be used as a Placement Test. It could be given to a student when he registers and, as a result of his own assessment, he could be allotted to the appropriate course. No time would be wasted in marking the test, and the student would have the feeling that he himself had decided at what level to study. Since this is not a do-or-die situation, the student would not be inclined to exaggeration in his self-assessment, for by doing so he would find himself in a class above his level and he would be the one to suffer. On the other hand, coming to a Self-Assessment Test "cold", (as is the case with our students who come to the university after three years of army service), the student might not be able to judge himself accurately.

In addition to its use as a Placement Test, a Self-Assessment Test would appear to have considerable diagnostic potential and, finally, once a reliable format and technique have been established it might be possible to use such a test as part of a Final Examination.

2 What Kind of Test?

The first problem was - what kind of test? Does one need a special kind of test or would a routine classroom test be suitable? A fairly obvious method is to ask the students - after they have finished a classroom test - to estimate the mark they think they should get on it; the test is then marked, returned to the students and the two marks compared. I presented a paper at the AILA Conference in Stuttgart in 1975 on some results of this method, but the whole operation is a fairly barren one, since very little can be learnt from it.

Another way is to use an X-point scale. In a Standardized Form

Method (as suggested by Oskarsson 1978), the student is asked to assess himself as follows: For Reading, "Imagine that someone with no knowledge of English comes to you with an English daily paper (e.g. The Times, The Guardian) and asks you to give as detailed an account as possible of what is on the front page. You have plenty of time but must not use a dictionary. How well would you manage the task?" (p. 40) The student is told to rate himself on an 11-point scale ranging from zero "I would not be able to say anything" to 10 - "I would be able to give a detailed account of everything on the page". This is extremely subjective and difficult. Another version suggested by Oskarsson is that of the "descriptive rating scale" in which each rank is spelled out; for example, the second highest rank in reading is "I understand everything or nearly everything written in the language within non-specialized fields. There may be words I do not understand in difficult texts"; and the one above the lowest is: "I understand the main point of a simple text and simple written directions for familiar things" (Oskarsson 1978:43). This is more satisfactory, but is still vague - what, for example, is a simple text?

I have recently been impressed by a method being used in Sweden today to test immigrant workers' knowledge of Swedish as a second language, in order to place them in an appropriate course (Von Elek 1982). This has two innovations: i) Students are not asked to carry out any language tasks, but they are asked whether they could carry out certain language tasks. For example:

a. Vocabulary. "Can you fill in the gap in the following sentence?"

   What _____ is that dress? It is red.

b. Reading Comprehension. "Do you understand this sentence?"

   Due to our well-organized pre- and post-natal care, the infant mortality is extremely low.

ii) The second innovation is that the students can give three possible answers:

1. Yes
2. I think so
3. No.

Von Elek found that the introduction of "I think so" was psychologically well-motivated, since it tended to prevent wild guessing and thus to make the "Yes" answer more reliable. In his research, the teacher went through parts of the test with each student. His provisional results showed that the students' self-judgements were highly reliable as far as the "Yes" results were concerned, the only major differences being with the grammar sub-section.

Although I did not feel so confident as Von Elek about the reliability of such a test as a Placement Test, it seemed to me to have great value for use during a language teaching course for three reasons:

1 Educational. Differences between students' self-assessment and

their actual achievement on the same questions should provide an opportunity for discussion between teacher and student as to the cause of such differences. It could result from lack of validity of the test, or from the student's unwillingness to face up to his weaknesses, or to his modesty or other factors which might not be revealed under normal circumstances. Consequently, I decided that immediately after the self-assessment test the students would be given a test with similar questions, but this time they would write in their answers and the two scores would be compared.

2 Diagnostic. Using my method makes it possible to pinpoint areas of weakness and/or uncertainty on the part of the student, and perhaps areas of weakness in the teaching. Once such areas have been discovered, ways of improving the teaching can be worked out. This, in fact, seems to be an improved form of Error Analysis.

3 Psychological. After being exposed to one or two self-assessment tests, the student will not only begin to have a much clearer idea of areas in which he is strong and those in which he is weak, but will also gain confidence as a result of this self-knowledge. This is something which is too often ignored.

Consequently, I have constructed self-assessment tests and achievement tests containing the same questions at different levels of difficulty, and used them with classes at different levels of proficiency. The aim was to investigate the reliability of such tests as Placement Tests, their potential value as diagnostic tests and the possibility of using them as part of an Achievement Test.

3 The Experiments

3.1 The Subjects

a. Students at the Hebrew University take a compulsory course in EFL with the main emphasis on Reading Comprehension. They study English for one or two years depending on their proficiency in English, and those studying for one year are divided into three levels. They may also study in an intensive summer course. Thus, we have four levels which I am calling Levels I, II, III & IV going from lowest to highest. Level I could be called Intermediate English and some of the material they study is edited and slightly simplified. Levels II, III & IV read only unsimplified material - Level II could be called Low Advanced, III Mid-Advanced and IV Advanced. It was expected that the differences in mean scores between levels on the same test would be about 10%. Students of Levels I, II & III have been tested - these are Groups P,O & R.

b. Another group of students tested were prospective English teachers studying at a Teachers' Seminary. It was expected that they would be approximately at Level III - this is Group S.

3.2 The Tests

Von Elek had six sub-sections, testing Vocabulary, Grammar, Listening, Comprehension, Speaking Ability, Reading Comprehension and

Writing Ability. In our university course, although the student is generally expected to answer in English, little emphasis is put on this skill, which is not taught, and it was decided not to test it. On the other hand, students are listening to English all the time, and so Listening Comprehension was tested. In the pilot study Writing Ability was tested, although it is not taught; however, it turned out to be irrelevant and was dropped. Thus, the final tests which will be reported on, Tests B and C, consisted of:

a. Listening Comprehension - two short texts, read out by the classroom teacher, with five questions on each - 10 items.

b. Vocabulary - opposites and fill-ins - 15 items.

c. Grammar - Fill-ins and correcting incorrect sentences - 10 items.

d. Reading Comprehension - a short passage with five questions and a longer one with 10 - 15 items.

It was planned that Tests B & C would be approximately suitable for Levels II & III, and that the average Level II student achieve about 60% on Test B. Here I set my sights too high. Level A & D tests are now being constructed.

3.3 Administration of the Tests

In the self-assessment tests the questions were phrased in the form of: "Do you understand...?" or "Do you know...?" or "Can you supply...?" and so on. The students marked their answers in one of the three columns on an Answer Sheet: "Yes"; "I think so"; "No". Only the answers in the "Yes" column were scored as correct answers.

The achievement test consisted of exactly the same questions, but here the instructions were to "Write the opposite of the word underlined" or "Fill in the gap" or "Rewrite the sentence in a way that you think is correct".

3.4 Time Limit

It was decided not to limit the students for time. It was estimated that about 30 minutes would be required. Some finished in less than that time, but no one needed more than about 35 minutes on the self-assessment. On the achievement tests the students required about five minutes more. Once the tests become more or less standardized, it might be necessary to fix a time limit.

4 Results

4.1 Results on Test C

Table 1 shows the results of Test C, given to students who were expected to be at Level III. Group S were First Year Seminary students on their first day of study, and Group R were First Year University students on the second lesson of their course. For them the tests could be considered as Placement Tests.

Table 1

Results on Test C

	N	Self-Assess. Mean	S.D.	Achievement Mean	S.D.	Difference Overestimate (+) Underestimate(-)
Group S (III)	14	29.8	6.6	21.1	8.9	+8.7
Group R (III)	32	27.9	9.1	19.6	7.0	+8.3

4.2 Results on Test B

Table 2 shows the results of Test B, which was given to Groups P & Q whilst they were in the midst of an intensive summer course, and to Group S two days after they had taken Test C.

Table 2

Results on Test B

	N	Self-Assess. Mean	S.D.	Achievement Mean	S.D.	Difference Overestimate (+) Underestimate (-)
Group P (I)	18	25.3	6.4	20.5	8.3	+4.8
Group Q (II)	19	26.8	10.9	24.7	9.9	+2.1
Group S (III)	14	27.6	8.0	31.0	8.2	-3.4

4.3 Summary of Results on Tests B & C

For ease of comparison, Table 3 summarizes the results of Tests B & C, with the figures given in percentages.

Table 3

Summary of the results of Tests B & C. Figures in percentages.

Test B

	N	Self-Assess.	Achievement	Difference
Group P (I)	18	51%	41%	+10%
Group Q (II)	19	54%	49%	+ 5%
Group S (III)	14	55%	62%	-7%

Test C

	N	Self-Assess.	Achievement	Difference
Group S (III)	14	60%	42%	+18%
Group R (III)	32	56%	39%	+17%

5 Discussion

Since there are so many variables here, there is no point in looking for statistical significance, and so we will look at certain aspects of the data from a number of different angles:

5.1 Accuracy of Self-Assessment

a. At the beginning of a course

It was (mistakenly) thought that Level III students (Groups R & S) would find Test B too easy and would be insulted by being given such an "easy" test, and so they were given Test C. As can be seen from Tables 1 & 3 their achievements were low, about 40%, but what is interesting is that not only did both groups over-assess themselves but that their over-assessments are remarkably similar: 17% and 18%. Two days later Test C was returned to Group S, the results were discussed and they were given Test B. Possibly the shock of Test C sobered them because, although their achievement was reasonable - 62% - this time they under-assessed. (Unfortunately, it was not possible to give Test B to Group R.) Admittedly, the tests had been given to Group S in the wrong order, but the results show that self-assessment tests at the beginning of a course are not likely to be reliable.

b. In the middle of a course

Groups P and Q overestimated themselves on Test B (Table 2) but not excessively. However, on the skill that they had been mainly

taught, i.e. Reading Comprehension, the differences between self-assessment and achievement were negligible - it was +2% for Group P, -2% for Group Q. This certainly suggests that students are capable of evaluating themselves reasonably accurately on skills that they have been taught.

c. In general

The results (Table 3) indicate that if a test is "difficult" (achievement less than 50%), students tend to over-assess their proficiency. The only case of under-assessment was with Group S on Test B, which was a "reasonable" test for them, since their achievement was 62%. However, it may well be that the "shock" of Test C contributed to this under-assessment. Whatever the reason, we are in a vicious circle here - we want the students to assess themse-ves at the beginning of a course, but the most reliable test for them is one suitable to their level of proficiency and we will not know that till they have taken a test! This means that using this kind of self-assessment as a Placement Test in a foreign language (as opposed to a second language, as in Von Elek's research) is of doubtful value.

5.2 Relationship between positive and negative self-assessment and actual achievement

Tables 4 and 5 give the results of two questions arranged to show what percentage of students actually gave acceptable/unacceptable answers as compared with what they had marked or not marked on their Answer Sheet.

Question 84 was in the Grammar Section of Test B under the rubric: "Can you see what is wrong with the sentence and correct the error?" Question 84 was: "The news about it everyone was talking turned out to be false."

Question 143 was in the Reading Comprehension Section of Test C. The reading passage contained the following sentence: "There are theorists today who, proceeding on the assumption that all children are naturally good, kind, intelligent, unselfish and discriminating, deprecate any discipline or direction from adults." Question 143 was: "Could you fill in the TWO missing words? "The argument of the'theorists' (line 5) are _____ the assumption that all children are naturally good." The phrase "based on", which was the expected answer, appeared in the next sentence.

Table 4

Results of answers to Question 84. Figures are in percentages.

	Gave an acceptable answer			Gave an unacceptable answer			Gave no written answer		
	P	Q	R	P	Q	R	P	Q	R
Said that they knew	-	16	36	39	-	7	5	-	-
Said that they did not know	5	5	36	28	47	14	22	31	7

Table 5

Results of answers to Question 143. Figures are in percentages.

	Gave an acceptable answer		Gave an unacceptable answer		Gave no written answer	
	R	S	R	S	R	S
Said that they knew	19	42	22	14	-	-
Said that they did not know	9	29	12	14	-	-
Marked nothing on the Answer Sheet	0	-	3	-	25	-

From Table 4 we see that Question 84 was too difficult for Group P (Level I) and not much easier for Group Q (Level II). However, in Group S (Level III), 72% of the students gave an acceptable answer (36 + 36), but only half of those stated beforehand that they knew the answer. Again in Table 5, we see that 71% (42 + 29) of Group S gave an acceptable answer and this time well over half of them stated beforehand that they knew the answer. Group R (also Level III) would appear to be much less confident of themselves than Group S, since 34% (9 + 25) of them wrote nothing on the Answer Sheet and 25% gave no written answer.

These figures call into question the reliability of the whole research. What is happening here is that the overall picture that we get from Table 3 is masked by opposing differences, since + and - differences cancel each other out; but we do see that Group 0 (Level II, whose mean score on the Achievement Test was 49%) show quite a stable picture - only one student (5% figure) estimating incorrectly. What has not been taken into account is the tendency of some people to over- and others to under-estimate themselves, irrespective of their language proficiency. This aspect was not investigated, but if such a "correction" could be introduced (by giving the student the appropriate psychological tests) it might clarify the picture. However, it would also make the whole procedure much too cumbersome.

5.3 Diagnostic Use

This is perhaps the most interesting aspect of the whole research. The incorrect rewritings of Question 84 are too many to list - here are a few examples of rewrites by students who claimed that they did know how to correct the sentence:

The news about everyone was talking turned out to be false.

The news about everyone was false.

The news was talking about turned out to be false.

The news about it which everyone was talking about turned out to be false.

Here are some of the unaccepted answers for Question 143 by students who claimed they did know the correct answer: against, to justify, to prove, proving that, about, keeping, believe that.

This is wonderful raw material for the teacher!

6 Conclusions

1 The kind of Self-Assessment Test, as used here, would seem to be unsatisfactory as a foreign language Placement Test for new students, although a series of graded tests presented in order of difficulty might produce more reliable results.

2 The most reliable area of Self-Assessment was with classes that had been studying for some time and with a skill, in this case ReadingComprehension, that they had been mainly taught. This produces the obvious conclusion that students can best evaluate themselves in the skills they have been taught!

3 Tests of self-assessment have great diagnostic value and they can be used with excellent effect during the progress of a language teaching course. They help both the student and the teacher to understand better the processes and strategies that the student is using.

4 Students who have been exposed to Self-Assessment Tests become more conscious of their language strategies and thus better at self-

assessment. Consequently, it seems feasible to use such tests as part of a Final Examination at the end of a language course, provided the students have had practice in them.

5 Finally, the type of Self-Assessment Test described here could be a valuable educational and psychological adjunct to the teacher's teaching, but it still has to be proved that it is a reliable method of self-assessment in a foreign language.

Bibliography

Berkoff, N.A. 1978. "Discrepancies between Student and Teacher Estimates of Language Proficiency". In Nickel, G. (ed). 1978. Language Testing. Papers from the Fourth World Congress of AILA. Stuttgart: Hochschulverlag.

Oskarsson, O. 1978. Approaches to Self-Assessment in Foreign Language Learning. Oxford: Pergamon Press.

Von Elek, Tibor. 1982. Tests of Swedish as a Second Language: An Experiment in Self-Assessment. Goteborg University: Language Teaching Research Center.

TESTING LISTENING COMPREHENSION: NOTIONS AND FUNCTIONS AS "DISCRETE POINTS" IN A COMMUNICATIVE CONTEXT

Marijke Boeijen (National Institute for Educational Measurement, Cito, Arnhem)

1 Introduction

In this paper two kinds of French tests will be presented that both originated from The Threshold-Level (T-Level) or Un Niveau-Seuil. We shall first have a short look at some of the concepts mentioned in the title of this paper; especially the concept of discrete point-testing with regard to Notions and Functions from the T-Level will be discussed. The concept of communicative context and the assessment of listening comprehension will be paid attention to. Specimens of the tests will be demonstrated and back-ground information will be given. Some psychometric data will also be presented.

The first type of tests are objective-based or criterion-referenced tests which have a formative purpose and are intended for the first years of the learning period. The learning objectives have been determined on the basis of the General Notions as described in the T-Level or Un Niveau-Seuil. The discrete-point approach is rather predominant here, whereas the communicative aspect is limited and could be characterized as pre-communicative.

The second type concerns proficiency-oriented tests with a summative function, at the end of the first period of circa 4 years. These tests cover parts of the Specific Notions on the one hand and of the Language-Functions on the other. The communicative aspect here is more evident and although the tests themselves deal with discrete points of the linguistic competence, these points are embedded in more holistic comprehension procedures.

2 Discrete-point testing and listening comprehension

When Cito started the construction of objective-based tests for listening comprehension the current tests in use for beginning students in foreign languages were of the discrete-point type: They usually measured isolated elements of the linguistic competence focusing on specific features of grammar. This procedure of breaking up the language into its smallest constituents cannot be used either in teaching or in testing listening comprehension. In a communicative skill as is listening comprehension, it is no use asking for the simple knowledge of constituting elements. Context in the largest sense of the word plays an indispensable part.

Still the need was felt to structure the complex listening-process into smaller units. It seemed possible to do so once the T-Level-model was published. In this model language-learning is defined in terms of practical communicative ability. The analysis made

on the basis also of extra-linguistic features (target-group, language activities, settings, roles, topics) led to an interesting categorization of the language-forms or exponents into General/ Specific Notions and Functions. (Van Ek 1976) Within the three main categories one can distinguish smaller units which can be seen as learning objectives. This kind of objectives seems appropriate to the teaching and testing of language in terms of skills. More particularly the General Notions (e.g. temporal, spatial, logical etc., see 5.1) constitute a finite domain. As far as comprehension is concerned, the frequent occurence of General Notions necessitates good mastery of their language-forms. Now these language-forms, centered round a common denominator that we will call a learning objective and which in fact is a subcategory of the General Notions, can be considered to be discrete points.

3 Notions and Functions in a communicative context

Measuring listening comprehension in a communicative context should not be equated with measuring communicative competence, although there is some relationship. The latter is a rather ambiguous concept open to different interpretations. (Canale, Swain 1980) Without going into a discussion on what aspects of human behaviour have to be included in its definition, we may safely say that communicative competence includes abilities which go beyond linguistic competence proper. However, any language-related communicative situation requires a certain level of linguistic competence. Even a if non-native speaker replaces his lack of command of the foreign language by gesticulation when he wants to express himself, he still has no control at all over the linguistic reaction of the respondent. His (receptive) ability in communicative situations therefore has to contain a good deal of linguistic competence. On the other hand his use of his communicative competence, when he has to understand a native speaker's words, can make up for deficiencies in his linguistic ability. So useful information may come from extra-linguistic factors such as the type of language-contact, from the setting, roles, topics, speech and/or listening intention. Therefore it would be artificial to separate the linguistic competence from the extra-linguistic factors that determine the communicative situation.

In the tests that are discussed here and that assess, only the listening or comprehension part of the communicative act, the testees are asked to demonstrate mastery of their linguistic competence. They are helped by the extra-linguistic guidance that comes from the specification of their role, of the topic, of the setting in that particular communicative act. And it is this guidance that we will consider to be the communicative context. It is meant to create or to substitute an authentic and intrinsical communicative intention, be it a speech intention or a listening one. Which Notions and Functions are used, is - partly - a consequence of the communicative intention.

4 How to assess listening comprehension

The communicative ability is often considered as the ability to convey information in the first place. The ability to receive information is given less attention. There are, of course, specific and exclusive listening situations. But apart from these, the importance of the receptive side in the communicative act is often underestimated. It is not the unpredictibility of the native speaker's reactions that tells strongly for the teaching and testing of listening comprehension. There are other arguments as well. The target-population of the tests we are discussing here consists of students in the initial phases of learning the language. Even if the most important objective of language learning was oral production as the main factor of communicative competence, there would be some arguments for training the receptive skills in the first years of the learning process in order to prepare in the most successful way the path towards oral proficiency. The so-called receptive semantic approach to language learning in the initial phase is gaining popularity. (Nord 1980, Postovsky in Winitz 1981)

However, listening and speaking are indeed very often integrated into a more complete communicative act. So we tried to take this fact into account by creating a communicative context and more particularly by guiding the specific listening intention as we explained before. For the tests this resulted in the simulation of (mini)-dialogues. The stimuli are supposed to be of such a low degree of difficulty that the testees could have produced them by themselves. The communicative need has to be immediately evident to them. The part to be tested is situated in the responses of the dialogues, the contents of which can be partly anticipated by the (easy) stimulus. The test-items focus on Notions or Functions.

5 The tests

Two types of tests will be described and demonstrated one after another.

5.1 The objective-based tests (General Notions)

For these tests 11 objectives were selected and formulated in terms of General Notions to be found in Un Niveau-Seuil (Coste 1976):

1 Interrogatives
2 Spatial Notions (location, relative position, distance, direction)
3 Temporal Notions (period, point of time)
4 Quantitative Notions (number, quantity)
5 Notions of degree
6 Notions of frequency
7 Deictic Notions (personal pronouns)
8 Deictic Notions (demonstrative pronouns/adjectives and possessive pronouns/adjectives
9 Deictic Notions (relative pronouns)

10 Contrastive Relations ((in)-equality, correspondence, contrast, comparison)
11 Logical Relations (con- and disjunction, in- and exclusion, cause, purpose, condition)

The overall goal of the tests is a formative one: to provide teachers and students (age-group 12-14 years) with a series of tests that can be used during the learning process to evaluate and to improve teaching-methods and learning-effects.

The learning-objectives are stated in behaviourial terms. General Notions from Un Niveau-Seuil were selected for three reasons. They occur in the curricula most frequently used for the beginning learners. They constitute a finite domain and they can reconciliate two opposite approaches: the discrete point and the communicative.

The language-forms belonging to a particular learning objective are, each, part of a test-unit. Each unit consists of a life-like situation or topic relevant to the target-group and is presented as a mini-dialogue. The stimulus - a question or a statement - is simple and is forced, as it were, upon the student as his own communicative need. An artificial procedure, it is true, but needed as a kind of advanced organizer of his listening intention.

The test-length varies from 5 to 20 units and the item-format that is used is multiple choice with 2 or 3 options preceded by a question, in the mother-tongue. Mastery of an objective is supposed to be reached at a test-score of about 80 per cent.

Objectives and their operationalisations into mini-dialogues and test-items have been pretested, revised and approved of in school-practice.

5.1.2 Some samples of objective-based tests

Learning objective 4: The testee must, in a short conversation, be able to understand notions that denote quantity.

Text on tape:	Test:
- Hé Michèle, combien d'argent as-tu dans ton portemonnaie?	How much money does Michèle's purse contain?
• Hélas, je n'ai pas d'argent du tout.	A There is no money in it at all. B There is still a lot in it.
- Les amis d'Henri étaient là, hier, au match de football?	Did Henry's friends go and watch the footballmatch?
• Il n'y avait que Paul. Les autres ne sont pas venus.	A No, only Paul did. B Yes, all of them did, except Paul.

Learning objective 10: The testee must, in a short conversation, be able to understand notions that denote comparison.

Text on tape:	Test:
- Tu as vu à la télé le match de tennis entre la France et l'Angleterre? ● Non, c'était bien? Est-ce que les Français ont perdu? - Oui, mais ils ont joué aussi bien que les Anglais.	What does the boy say about the Frenchmen's play? A They played better than the English. B They played just as well as the English. C They did not play as well as the English.
- Tu as beaucoup d'heures de français, cette année. ● Trois heures seulement. Ça fait une heure de moins que l'année dernière.	What does the girl say about the number of lessons in French this year? A There is one lesson per week more than last year. B There is one lesson per week less than last year. C There are just as many lessons per week as last year.

The underlined words are the language-forms belonging to the particular learning objective that is being assessed here.

5.1.3 Discrete point versus context

A major problem when constructing these tests is the question of context: does the context or the specific trait to be measured (time, comparison etc.) account for differences in scores? An answer to this question may be obtained from a comparison between each test and its parallel-test. For each learning objective two different tests have been constructed: test and parallel-test. Test and parallel-test contain the same elements of the learning objective, but differ in context and topic. General construction procedures have been the same.

Table 1 shows the scores on the test and parallel-test (pre-test version) dealing with the objective: logical relations. It is rather a long test and the figures may be considered representative of the test battery as a whole.

Table 1

Scores on the General Notions-test: logical relations, and on its parallel test

itemno.	$\bar{p}$	parallel $\bar{p}$	itemno.	$\bar{p}$	parallel $\bar{p}$
1	.97	.95	11	.86	.89
2	.91	.83	12*	.73	.96
3	.88	.88	13	.66	.78
4	.95	.81	14*	.92	.72
5	.88	.91	15*	.92	.51
6	.37	.38	16	.92	.96
7*	.90	.60	17	.97	.93
8	.84	.85	18	.86	.82
9	.84	.86	19	.96	.89
10	.88	.80			

Given the small differences in test scores between parallel-items one may conclude that in general, the global understanding of the contents is not such a strong factor that it outweighs the discrete point to be measured. In the cases preceded in the table by * disparallellism can be retraced to a lack or a surplus of redundancy in the context. In most cases redundancy has been reduced.

A confirmation of the non-global character of this way of testing may be found in the correlations between the separate tests each assessing different learning objectives, fluctuating from .27 to .55. These correlations show rather great differences. They are, as expected, generally lower than those between the subtests of the usual listening comprehension tests of Cito, that all aim at the assessment of global listening comprehension (r = circa .60).

5.2 Summative tests (Specific Notions and Functions)

We will now have a look at the second type of tests, the summative ones, that have been developed as subtests in a test battery for the lowest proficiency-level of Dutch secondary education. The whole of the battery was designed to assess the achievement of individual students after some 4 years of French.

5.2.1 Some history

The overall objective for listening comprehension-assessment at proficiency-level has been formulated as follows:

"the ability to understand the foreign language, spontaneously spoken by educated native speakers at normal, conversational

speed and excluding extremely informal elements, lexical and syntactic elements that are incomprehensible to less educated native speakers and also excluding topics requiring specialistic knowledge". (Groot 1975)

For the assessment of this ability we have since the early 70s used one type of text: the interview subdivided into smaller parts, and one item format: the multiple choice-question consisting of one question and 3 answers, each presenting a one-phrase-summary of the global contents of each part. The items are formulated in the target-language.

Problems arose with respect to the finding of spontaneously spoken language (interviews) at the level under consideration. Other problems are connected with the exclusive implemention of the overall objective by interviews. There are far more listening situations and tasks than only those in interviews; and one can think of other item-formats as well. Validation research undertaken at our institute showed the constraints that the exclusive use of our multiple choice-technique imposed on the language material. (Van den Nieuwenhof 1979, De Jong 1982a/b) Diversification was to be found in both listening tasks/situations and question-types in order to enhance the validity of our tests. We tried to attain a greater variety on the basis of Un Niveau-Seuil, since we are dealing with the lowest level in our educational system. After having listed the possible listening situations/tasks, the topics and the item-formats, we made an enquiry among teachers about their and their students' preferences. These turned out to be the (traditional) interview, the dialogue or discussion, real life-like situations and samples of radio broadcast. As to item-format there was no real consensus.

5.2.2 Characteristics of the new tests

Three new tests such as students and teachers had shown preference for, were developed and tried out:

Information or messages passed on through the existing channels such as radio-broadcasts, preserving the characteristics of tone and language, and on topics suitable for the target-group.
The test requires testees to grasp the global contents of the topic (open listening intention) on the one hand and to understand detailed information like time, date, prices etc. (guided listening intention) on the other. The item-format is the multiple choice-question: 3 statements in the mother-tongue.

An every-day like situation in which the roles of the partners in the conversation are formal ones.
The test measures mainly mastering of Specific Notions (in the example given here, those having to do with buying clothes in a shop) as they are current with respect to a particular topic. The listening intention is therefore strongly guided. The item-format is multiple choice: 2 photographs, one of which represents the interlocutor's words.

A discussion that is characterized by the informal relations

between the partners. The topic is, again, supposed to be relevant to the target-group. The test is meant to assess the mastery of the Language-Functions that express emotions. The listening intention is guided towards the emotional or attitudinal aspects of the words spoken by each of the partners. The student is, as it were, overhearing a conversation, between youngsters of the same age-groupe. The item-format is multiple choice: 3 statements in the mother-tongue.

These new tests have been integrated in a test battery which also includes a traditional interview.

5.2.3 Some examples from the every-day like situation test: shopping (Specific Notions)

We have selected an example from this particular test because it corresponds most with the discrete point approach in a communicative context. The listening intention is strongly guided. Testees are asked to identify themselves with the client-part and to focus, as far as the test is concerned, on the answers of the shop assistant. These become clear in one out of two photographs.

La cliente: - Le pantalon bleu dans la vitrine, je peux le voir s.v.p.?
La vendeuse: • Mais bien sûr; venez avec moi, si vous voulez; vous allez me montrer lequel, car il y a deux pantalons bleus dans la vitrine.

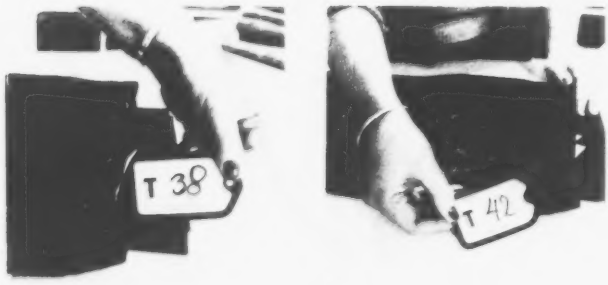
A

B

La cliente: - Je peux essayer ce pantalon en taille 40, s.v.p.?
La vendeuse: • Bien sûr. Les cabines d'essayage sont derrière vous, regardez ... là-bas.

A B

La cliente: - Ça ne va pas.
La vendeuse: • Ce modèle taille très petit. Voilà, essayez la taille au-dessus. Le 42 vous ira certainement mieux.



A B

5.2.4 Problems

Photographs are used in this test for various reasons. We wanted to diversify our item-format staying however within the range of the closed question-types. Photographs have the advantage of avoiding language-problems like having to read statements either in the target-language or in the mother-tongue. At the same time they give some of the extra-linguistic support a non-native speaker would have made use of in the real situation.

In table 2 the data of this test are given in relation to those of the other tests of the battery in order to enable us to make a comparison. The test we are discussing here is half of the test that is labelled A. (A2)

Table 2

Results on the Specific Notions-test compared to the rest of the test battery

	n	$\bar{x}$	$\bar{p}$	SD	r (KR20) tt
A1* A2	23(23)	19.35(17.85)	.84(.78)	2.17(2.13)	.46(.25)
B*	12(12)	10.14(9.20)	.85(.77)	1.66(1.63)	.50(.36)
C*	15(15)	12.57(10.60)	.84(.77)	1.98(1.94)	.53(.38)
A B C	50(50)	42.06(38.67)	.84(.77)	4.41(4.33)	.70(.60)

Note: N = 972(107)
* = the separate parts of the test (the subtests)
A1 = radio information, A2 = Specific Notions (shopping)
B = interview, C = discussion containing emotional and attitudinal Functions

Test A proved to be less reliable than our usual tests of which B is a representative specimen. Test C shows more or less the outcome of our standard tests which is not very surprising since test C is in its form closer to the normal interview-type.

Reliability, however, is a problematic concept in the outcome of this specific test battery. Although test A, B and C are supposed to measure testees' listening comprehension at the lowest level of the Dutch secondary educational system, this battery is always taken together with a set of tests meant for one level higher. Students having followed the same curriculum can take examinations on two different levels. As to listening comprehension, the decision procedures to assign students to one level or to the other, are based on the assessment of all students' listening comprehension on both levels. This means that a large proportion of students belonging to the next higher level distort the data shown in table 2. There is no other means of distinguishing in advance between the two levels in the population. Finally about 20 per cent of the population in table 2 is to be found on the lower level we aim at in the tests discussed here. In table 2 the figures of the actual target-populations for which these tests have been developed, are given in brackets.

The fact that reliability is even lower here than for the whole population is not surprising since the normal distribution cannot be expected. After all, students belonging to this group (in brackets) are the lowest scoring testees of the whole population of table 2. Nevertheless the test battery developed for the lowest level in combination with the tests for the higher level comply with the objective to make a distinction between the two levels with a sufficient degree of certainty ($p \leq .01$) (see Appendix)

Yet test A remains a problem, especially the photo-section. It hardly contributes to the assessment of testees' listening comprehension. The item-format might account for weakness. At this moment further validation research is undertaken on this particular issue. The intrinsic reliability of this test will be measured by a comparison with two other item-format for the same contents (text).

6 Conclusion

We have discussed and shown examples of two different types of listening comprehension tests based on different Notions of the T-Level and Un Niveau-Seuil.

The first type, learning objective-based tests, has been developed to enable teachers and students, in the initial phase of the learning period, to find out about mastery of discrete points (General Notions) in a meaningful context. We think that we have succeeded in reconciliating, to a certain extent, two seemingly opposite approaches.

As to the second type, summative listening tests we have shown our efforts to integrate an other kind of discrete-point testing (Specific Notions) into a broader communicative context, together with a variety in item-formats of the closed question-type. As far as the photo-test is concerned, we encountered problems of reliability, but we think this kind of tests might be an enrichment of our test battery as regards both contents and item-format. Further development of this technique together with further validation research seems worthwile.

Appendix (with acknowledgement to my colleague John de Jong)

For this particular school type final examinations are administered at two different levels. Most pupils aim at the higher level and in fact 80 per cent succeed in taking this higher level. However, decisions on the level at which pupils graduate , are based on their achievements on tests they have already taken. This is why all pupils of this school type partake in tests of both levels. Depending on their results the final decision on their graduation level is taken and consequently only the results of those tests that are meant to assess that particular level are taken into account. In this situation it is essential that tests for each of the levels should clearly differ in required ability. Data on classical test analysis of both tests are presented in table 3.

Table 3

Results of one population on two tests of different levels

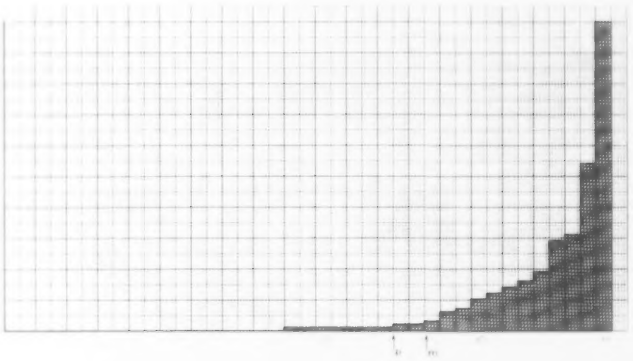
	higher level tests	lower level tests
n	38	50
p	.72	.84
SD	4.41	4.41
r (KR20) tt	.69	.70

Note: N = 972

Product-moment-correlation between the two tests is .73. Classical test analysis including p-m-correlation, however, gives a global evaluation of test quality only and does not take into account discriminating power at different sub-levels and consequently neglects differences in standard error at different ability levels. Of course, one would expect higher scores for alle pupils on the easier (i.e. lower level) test. But pupils who succeed on all items in the higher level test cannot be expected to obtain an even higher proportion of right answers on the lower level test. To illustrate this point, figure 1 presents - at each different score in the higher level test - violations against the expectation of a higher score on the lower level test. Violations are defined here as pupils obtaining a higher or equal proportion of right answers on the higher level test. The violations are expressed as a proportion of the frequency distribution: the total number of pupils obtaining that particular score. In the critical area below the passing score less than .01 violation can be observed. It is only well above the mean score that violations constitute a considerable proportion of the observed scores and all pupils with a maximum score on the higher level test obtained an equal high score on the easier test. Misclassifications of pupils appear to occur less than indices from classical test analysis would suggest.

Figure 1

Violations - as proportion of observed scores - (y-axis) versus score on higher level test (x-axis)



Bibliography

Canale, M., Swain, M. 1980. "Theoretical basis of communicative approaches to second language teaching and testing." Applied Linguistics 1: 1-47.

Coste, D. et al. 1976. Un niveaul-Seuil. Strasbourg: Council of Europe.

Ek, J.A. van, 1976. The threshold level for modern language learning in schools. Strasbourg: Council of Europe.

Groot, P.J.M. 1975. "Testing communicative competence in listening comprehension." In Jones, R. and Spolsky, B. (eds.) 1975. Testing language proficiency. Arlington, Virginia: Center for Applied Linguistics.

Jong, J.H.A.L. de, 1982a. Testing foreign language listening comprehension. Paper presented at ISET 5, Stirling Scotland.

Jong, J.H.A.L. de, Nieuwenhof, H.W.M. van den, 1982b. Een experimentele luistervaardigheidstoets. Arnhem: Cito.

Nieuwenhof, H.W.M. van den, Jong, J.H.A.L. de, et al. 1979. Validatie van luistertoetsen moderne vreemde talen. Arnhem: Cito.

Nord, J.R. 1980. "Developing listening fluency before speaking: an alternative paradigm." System 8: 1-22.

Postovsky, V.A. 1981. "The priority of aural comprehension in the language acquisition process." In Winitz, H. (ed.) The comprehension approach.

Winitz, H. (ed.) 1981. The comprehension approach to foreign language instruction. Rowley, Massachussets: Newbury House.

CAN GRAMMATICAL CORRECTNESS AND COMMUNICATION BE TESTED SIMULTANEOUSLY?

Barbara Falk (College of Education, Utrecht, The Netherlands)

1 Background

The purpose of this presentation was to invite discussion and advice as to the assessment of students' knowledge and use of the grammar of English by means of both written and oral tests of a communicative nature. These students are to become teachers of English as a foreign language in the Netherlands, where standards of correctness have traditionally been high, almost to a fault. Yet, as more and more emphasis has been placed on "communication" in the foreign language, it seems that standards of correctness become increasingly more difficult to maintain, particularly as the interlanguage we fondly call "Dunglish" is most intelligible to native speakers of English (becoming, as interlanguages are wont to do, increasingly more intelligible as the native speakers become more experienced in listening to it, or on occasion, use it themselves.)

While retaining a primary interest in "the message", and accepting the fact that the effect of error will in our case be more psychological and sociological than actually distortive of content, we continue to strive for a mastery of standard English which approaches the "educated native speaker's" as closely as possible.

2 The tests

An attempt to measure this mastery is made once before the students are admitted to the school practice phase of their training program and again, generally two years later, before they are finally certified to teach. One written and one oral test are given. The written test (See Appendix 3) is based upon one ore more sometimes marginally edited authentic texts, usually involving dialogue, in which selected blanks are to be filled with correct grammatical forms. Cues range from zero to multiple choice selection (See Appendix 2). The oral test takes the form of either a traditional one-to-one interview, with a second staff member present as evaluator, or of a discussion test, which is video-recorded and subsequently assessed by two staff members working simultaneously but independently. Global assessment criteria are used for both forms of oral test. Disagreement is resolved by averaging, argument or arbitration.

3 Problems

The following difficulties continue to plague us:

3.1 The written test, like any Cloze or Cloze-type test, requires reading comprehension at text level as well as grammatical operations at sentence or phrase level, and is as such a "pragmatic test", the only kind of language test, in Oller's view, worthy of the name.

Nonetheless, we question the fairness, c.q. validity of the (mis-)information being obtained when grammatical errors may be attributable to misinterpretation of the text rather than misconceptions about the grammar rules of English. If we were focusing on reading comprehension, we might wonder about the validity of penalizing errors, such as the contrast between much and many, which would not display misinterpretation of the text, but rather lack of mastery of a formal distinction. Neither staff nor students are anxious to return to strictly discrete point grammar testing, finding the present test form relevant and meaningful, as well as generally interesting to do. How seriously should we be taking our misgivings about the "impurity" of our testing procedure?

3.2 In the oral tests we continue to maintain that "effective communication" is the main, sine qua non criterion for success. We do not, however, have a clear or specific and at the same time manageable definition of what this is. We are able to count errors and note "strikingly good formulations" to the candidate's benefit, but we cannot quantify "communication." We find ourselves more sensitive to errors when overall content or sophistication of argument is poor, and vice versa. Can we manageably quantify and set a norm with any semblance of objectivity involving both "number and seriousness of errors" and "quality and effectiveness of communication"?

3.3 Even if we succeed in overcoming the above problem, we would still be faced with the need to elicit a wider range of structures than has often proved to be sufficient in the topical interview or discussion test. Upon analysis we are often amazed to find some of the most satisfactory performances have been given with recourse to little more than the Simple Present Tense, or even a rich collection of BE+ predicate noun/adjectives. We are certainly working on developing flexible elicitation techniques with which the interviewer can lead (force?) the candidate into using particular structures without making the interview totally pre-digested or largely predictable. One of the main problems we have is how to avoid using the target structure ourselves, so that, strictly speaking, we would have to accept an elliptical answer from the candidate, e.g.

Q: What would you do in Reagan's place?

A: Shoot all the missiles to the sun.

Another problem is suspecting avoidance, but being unable to penalize it when the "job gets done", e.g.

Q: What advice would you give this woman? (attempt to elicit should/ought to)

A: None. It's her problem.

We would like to establish, on a better basis than our mere intuition, how many and which structures we must succeed in eliciting (correctly or incorrectly realized) before the test can be considered to have covered a sufficient range of the grammar. How many of the structures must minimally be correct? Whould we distinguish between serious, minor, insignificant errors? What criteria do we use for deciding how "bad" something is? (See Appendix 1 for our present working list

100 mm
A5
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
1.0 mm
1.5 mm
2.0 mm
ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz
1234567890
PG

of grammar points to be focused on).

4. Discussion, Advice, Comments

The discussion which followed both in the limited official time remaining and over lunch, coffee, tea and drinks, resulted in the following useful comments and pointers:

1) If we are interested in correctness, we must test for it, and forget about matters like seriousness of content and sophistication of argument, or at least exclude differences among candidates with respect to the difficulty level of content.

2) If we are interested in spontaneously correct speech, the written test is probably superfluous, or could be used only as a back-up check of cognitive compentence for those who fail the oral test. This of course assumes effective elicitation of a broad range of structures in the oral test.

3) One might be able to improve the efficiency of the testing program by using the written test as a screening device. Only those who passed the written test would be interviewed. If ultimately was shown that the written test is a good predictor of success in the interview, as far as grammar is concerned in any case, then the costly interview procedure could either be dropped or used in a limited form as a formal test of pronunciation.

4) We should not "worry too much" about the role of reading comprehension in our grammar test. But if we really feel "so guilty about it" we could select texts for readability with a comparable group prior to working them into tests.

5) It certainly is difficult to determine what "effective communication" is. The age and experience of participants in a discussion affects judgement as does the age and experience of the person doing the judging, particularly as this is based on one's experience with certain native language and educational backgrounds. The video-taped discussion which was praised as "impressively good" by one conference participant was called "boring non-communication" by another. The discussion form may seem promising in its efficiency, but will never give us the information we're looking for with respect to usage. We are advised to keep it as a diagnostic exercise, but not to base serious decisions on it.

5. Conclusions

By and large, the author concludes the following:

1) It would be well worth our while to analyze our tests closely. There seems to be a discrepancy in what's measured by the oral and by the written test, but is it smaller than we think? Could we indeed achieve our purpose with only an oral test?

2) We should definitely continue our efforts to develop a flexible and yet highly structure interview procedure with which we can elicit a wide range of critical structures.

3) We will have to be content with our subjective consensus as to quantifiable norms.

The author hopes she can report on further developments with respect to these points at a future conference.

Appendix 1

WORKING LIST OF GRAMMAR STRUCTURES TO BE ELICITED (WITH EXAMPLES OF TYPICAL "DUNGLISH" ERRORS)

1. contrast simple vs progressive form

   X They are having a scheme to fill in.
   X I have been doing the phonetics course three times.
   X He was saying thank you when I gave him a present.
   X My little brother plays in the garden right now.

2. simple past with adverbials of past ime or clear past context

   X Many children have died in the Second World War.
   X I have handed in my homework last week.
   X Shakespeare has not written some of his own plays.

3. Present perfect (progressive) to express up to now idea

   X I live here since 1964.
   X Jack is playing in the top team for 10 years now.

4. infinitive after DO auxiliary

   X We didn't knew him.
   X Does he ever comes on time?

5. Future marker to express future time in main or independent clauses (except for appropriate use of present tense forms)

   X Theo is ill until next Wednesday.
   X I give this to you in a minute.

6. No will/would in temporal or conditional clauses of future time

   X If I'll see him tomorrow, ...
   X When you will die, ...

7. Inversion in questions rather than rising intonation on statement form, unless this is the appropriate form for the context (expressing surprise, doubt, etc.)

   X That is a good book you're reading?

8. some/any (-one, -thing, -where)

   X I know I've seen you anywhere.
   X There aren't some eggs left.

9. much/many

X Not much people read good newspapers.

10. any/every (-one, -where, -thing)

X It's so easy everyone can do it.
X You can come every day in the week, but only once.

11. who/which

X The people which I mean...
X The rivers who divide the country...

12. adverb/adjective

X He works terrible slow.
X She's very good built.

13. have to/should, ought to; had to/should have, ought to have

X This is the homework I had to hand in last week.
X If you think I have to do the course again, I'll accept your advice.

14. for/since

X He's been here since 3 days.

15. comparison of adjectives

X It's the worsest meal I've ever had.
X I'm doing worser and worser at English.
X That's the most nicest thing you've ever said.

16. limited use of <u>there</u> construction

X There are flying seven planes a day to New York.

17. this, that/these, those

X This people...
X These man...

18. It is/they are

X It is the dogs that pollute the park for the children.

19. infinitive constructions

X I want that you do this for me.

20. various frequently occurring prepositions

X <u>on</u> school , react <u>on</u>, discriminate somebody, etc.

Appendix 2

GRAMMAR TEST INSTRUCTIONS

There are FOUR main types of blank-filling cues:

I. No cue given in brackets. Blank only:

Write any grammatical word or phrase which completes the context. Grammatical words are prepositions, conjunctions, pronouns, auxiliaries and the like, NOT lexical words like small, dog, jump, green, etc.

Example: The new tax law discriminates ..... working women.

Answer: against

II. Brackets containing one word or group of words separated by commas (,):

a. Write down the correct form and order of the words given. Use everything given in the brackets, along with any helping verb (form of DO, BE, HAVE, WILL/WOULD). Do not use "modals" (see below) on this kind of item. In some items you may have to add a preposition (OF, TO, FOR, WITH, etc.) to complete the verb phrase.

Example: .....(you, always, smoke) when you're alone?

Answer: Do you always smoke

Example: Do you object .....(come) early tomorrow?

Answer: to coming

Do not make any additions to the words given other than the necessary auxiliaries or prepositions.

b. If AND ONLY IF you see an (M) in an item, use a "modal" (a modal auxiliary such as MUST, MAY, SHOULD, etc. or an expression of modality such as BE GOING TO, HAVE TO, BE ALLOWED TO, MANAGE TO, etc.) Do not use WILL/WOULD as "modals" on this test. Remember, do NOT use a "modal" unless there is an (M) in the item.

Example: I don't really remember who borrowed my book. It .....(M, be) Sam.

Answer: may (might, could) have been

III. A series of words separated by slashes (/):

Copy the correct choice on your answer sheet. Only one possibility is correct. -- means nothing is correct.

Example: John works evenings as .....(a/an/the/--) ambulance driver.

Answer: an

IV. Brackets containing one word either an adjective or adverb:

Write down the correct form of the word given.

Example: We walked .....(slow) away from the plane.

Answer: slowly

Note: At first sight these instructions seem overwhelmingly complex. In practice however, they cause virtually no difficulty. Most of the items in the test rather speak for themselves. We always do at least one practice test with a group of students before they encounter their first "real" grammar test, so that students in fact are fully familiar with the blank-filling variations. The detailed instructions are handed out with the test paper as a back-up reference.

Appendix 3

SAMPLE SECTION OF GRAMMAR TEST (104 items in total)

CHILDLESS BY CHOICE
(adapted from Newsweek magazine, January 21, 1980)

In Framingham, Mass., music teacher Leni Blumenthal and her husband Stan have decided not to have children because they fear a child could strain their eight-year-old marriage. Linda McGrew, who is married and runs her own vineyard in northern California, has rejected motherhood because she feels a woman's intelligence is better served "if she's not constantly driving the kids to ballet, gymnastics and piano lessons." Oregon bookkeeper Marilyn Cripe is a dedicated nonmother because she already has more things to do than she can handle. "I have hobbies and interests that I barely have time for now," she says.

..1.. (support) by the ..2..(women/women's) movement and the ease of birth control, more and more well-educated, career-oriented women are remaining childless by choice these days. Researchers 3....(study,now) the psychological causes and effects of rejecting parenthood, and a 2000-member support group for childless couples ..4..(spring) up, called the National Alliance for Optional Parenthood. The decision to reject motherhood is not an easy ..5..(one/--). "A woman ..6..(choose) not to have children ...7..(M, be) very ...8.(strong)," says Wayne A. Myers, a psychiatrist ..9..New York Hospital. "It's a rejection 10..(be) like her mother. She ..11.(wonder, also)..12. she's a selfish person and ..13.(M, worry) that she's not feminine."

For some ..14.(one/ones/--) the decision is ..15.(relative) easy to make and live with. The demands of working on a daily newspaper made it quite ..16.(impossible) for me to consider ..17..(have) children," says Charlotte Curtis, associate editor of the New York Times...etc.

ADVANCE PREDICTION OF DIFFICULTY WITH C-TESTS

Christine Klein-Braley (University of Duisburg)

Perhaps the major criticism levelled at my study of cloze tests (Klein-Braley 1981a) by my linguistic assessor was that I had failed to go sufficiently deeply into the question of text selection. This was a justified criticism. However it is not a criticism that applies exclusively to my work. In my search through the relevant literature I was unable to find anywhere guidelines for text selection. Even Oller, who devotes a whole chapter of his Language tests at school (1979), to cloze tests fails to provide the necessary help. By implication he suggests indeed that any and every text will serve the purpose. Yet when I began to develop cloze tests myself I found that there were in fact many more unsuitable texts in the world than suitable ones. It seemed to me at the time that there were two aspects to this problem. One was the question of text content, and the other was the question of text difficulty.

On the content side, it seems unfair to use texts which demand specialised knowledge from the examinee, particularly since in the classical cloze test only one long test is involved. However it is difficult to decide where general knowledge stops and specialised knowledge begins. Literary texts seem unsuitable, and older literary texts (Shakespeare?) seem doubly unsuitable, since in these cases an aesthetic and possibly an archaic factor is at work. Humorous and highly colloquial texts also cause problems because of their special linguistic characteristics. What is then left tends to be the contemporary, non-literary, discursive, argumentative or narrative text with no particular idiosyncracies or special terminology. What could, in brief, be called an unmarked text.

So far as text difficulty is concerned, the major assumption seems to be that the teacher or tester developing the test can select 'suitably difficult' texts for his examinees. Again, no additional guidelines are given, although Oller does claim that teacher judgements of difficulty are vastly superior to other more objective techniques, such as readability formulae, for example.

So, in cloze tests, we have a test form for which reliability and validity are claimed to be generic, while at the same time the basis for the tests, namely the texts, are a matter of subjective decisions.

Since the C-Test is a development of the cloze test, or, as we would prefer to formulate it, a different and more satisfactory operationalisation of the construct, we too are confronted with questions of text and test content and with questions of text and test difficulty.

We have not, so far, made any progress in the question of text

content. We do submit the texts selected by one of us to other judges and ask them whether they find the texts sufficiently neutral or unmarked to be used in a C-Test. But other than this we have not been able to define the quality of unmarkedness more exactly. Where we do feel we have made progress is in the area of text difficulty.

The C-Test is a norm-oriented test, and as such, aims at a medium level of difficulty on average. The mean score should be around 50% in order to ensure maximum differentiation between subjects. Inside the C-Test, however, it is preferable to order the texts in ascending difficulty, starting with an icebreaker and finishing with a really tough item. If necessary we can afford to to let the mean difficulty slide up to 60%: C-Tests have been shown to have high reliabilities even when they were much too difficult or too easy for the subjects involved and we feel it is important forthe icebreaker to be so simple that every test subject understands exactly what the C-Principle demands from him or her.

Since the subjects involved so far in C-Tests have ranged from 7-years-olds to adults, achieving mediumdifficulty in the tests for every subject group means that we need texts with a variety of difficulty levels, and that subjective jugdements of appropriacy are impossible to to make. In addition we have been developing tests for three different groups of addressees - L1 learners, L2 learners and LF learners. Moreover we have been working with two different languages - English and German - and this complicates matters rather.

For L1 groups we have been surprisingly successful in constructing appropriate tests, but this is because we cheated. Here we used other people's judgements and chose the texts from readers for the target groups, and combined them with texts from readers aimed at older and younger age groups. Obviously the authors and editors of such text books had some sort of procedure, either objective or subjective, which helped them to determine the adequacy of the texts. The area in which we had major problems was the LF advanced groups in Duisburg - learners of English - and I succeeded in developing appropriately difficult tests for them only by a process of trial and error. However, this is a time-consuming procedure, particularly when one is working inside a system where sufficiently large groups present themselves for testing only once a year. We felt that there had to be a better method than subjective judgements.

My first step was to discover what had been tried out so far. These methods can be grouped into text-oriented and subject-oriented techniques. The subject-oriented techniques include ratings, cloze tests and multiple-choice tests, and they all involve confronting the subject and the text in some way. He must rate the texts, or complete cloze tests constructed from the texts, or answer questions on the texts. Ratings are notoriously problematical (cf. Bartenwerfer & Raatz 1979) and in a small pilot study I found that neither German university students of English nor their academic teachers were particularly unanimous about their judgements of texts with regard to difficulty. Cloze tests were non-starters because both Charles Alderson (1979) and I had shown that the n cloze tests that can be con-

structed from a text using nth word deletion are all different tests. In addition I had been able to demonstrate that the difficulty level of a cloze test correlates with the number of structure words deleted. But since the different starting points delete different proportions of structure words, there is no such thing as the 'cloze difficulty' of a text. Bensoussan (1982) was able to show that multiple-choice comprehension questions have a difficulty level of their own, which is not necessarily related to the difficulty level of the text they are attached to, and in an - as yet unpublished - paper presented to the ACROLT Symposium in Israel in 1983 she was able to pin down the technical changes which made the items easier or more difficult.

The text-oriented methods all make use of statistical characteristics of text to predict difficulty levels for specific groups od addressees in advance. In general they take the form of multiple regression equations where two or more indicators are used to predict text difficulty in advance without ever confronting the text and the reader. However there are two major snags. The first is that they produce entirely different predictions. In a large-scale study Stokes (1980) showed that there was no one pair of formulae which consistently predicted the same difficulty levels. In an exploratory study (Klein-Braley 1981b) I was able to showthat formulae predictions also failed to agree with ratings or with cloze test results. The second snag is that the validation criteria for the text-oriented methods were in all cases the subject-oriented methods whose short-comings I have just reported.

One of the problems in readability research is that the construct of readability simply has not been adequately defined. This is shown, for example, by the fact that the readability researchers are not even unanimous about such matters as the percentage of correct answers in a multiple-choice reading test which they will consider as adequate for demonstrating that comprehension has taken place. Some insist on 50% right answers, others on 75%, others on 90%.

Another approach to readability would be to take the findings from text linguistics and apply these to texts. Unfortunately, as Pugh pointed out in 1981, thes methods are still very much in their infancy, and so far they do not appear to be sufficiently reliable and replicable for use by the layman. Furthermore the question of text difficulty has not been central to such research.

For this reason the readability research seemed to us to be much more relevant to our problem. These approaches had the advantage of being relatively simple, even though we were not satisfied with the results produced. However our problem, though related, was not identical to that of readability researchers. Readability does not really interest us at all. Our aim is much more modest. All we want to be able to do is to predict in advance how difficult a given text is likely to be for a specific group of addressees. Our criterion is quite simply the mean score obtained by that group on a text multilated in accordance with the C-Principle.

We therefore decided to use the same approach as that of the readability researchers, i.e. to attempt to relate the statistical characteristics of the texts and the empirically determined C-Test difficulties. A survey of the relevant literature provided a large number of indices which had been used in previous research. We immediately discarded all approaches based on word lists, since these are available only for English, and even here apply only to L1 groups. For German there are no such word lists, and a simple translation of the English lists would not necessarily produce equivalent procedures. We also chose to disregard any indices which demand too much human attention. We had a very small personal computer at our disposal, and it seemed fitting that such things as counting and matching should be done by machine rather than by woman. Since almost all regression equations available include one index which is in some way related to syntactical complexity - for instance sentence length - and one which is related to the range of vocabulary, we ensured that such indices were included in our analysis. The type-token ratio seemed particularly attractive as a vocabulary index since it is not language-bound, can be easily worked out by the computer, and has been advocated by respected scholars working in the field of linguostatistics (cf. Herdan 1966).

Our procedure was quite simple. We first administered a number of texts to various sample groups and determined the C-Test difficulties for the individual texts. At the same time we calculated a variety of different indices for the texts. We decided to begin with a two predictor regression equation. In order to produce this we next correlated the text indices with the empirically determined C-Test difficulties (P Values). The two indices which had the highest correlations with the C-Test difficulties were then combined in a multiple regression, assuming for the moment that there was no suppressor effect operating.

There are many more Germans and learners of German available for testing in Germany than there are native speakers and learners of other languages. For this reason our German results are much better substantiated and complete than our results for native speakers and learners of other languages. I would therefore like to begin with the German groups.

Last year in Duisburg we were able to present a regression equation for German L1 speakers. The two predictors were the type-token ratio and the number of sentences in the text. This regression equation was derived for third grade pupils on the basis of tests administered to fairly large groups. However we later decided that the number of sentences in a text was very much the result of a subjective decision by us as to where the logical unit 'text' should end. Often one can manipulate these values by simply adding or subtracting a sentence. Since manipulation was not the point of the exercise we decided to use an index which correlated very high with the number of sentences in the text, namely the average length of the sentences in words. An average tends to be less susceptible to minor changes in text length.

Using empirical data for 12 texts and a total of 201 pupils in the third grade we calculated a regression equation for prediction of empirical difficulty in this group using these two indices:

Y' = 23P - 203 TTR - 1.9 XSL(W)
(Multiple R = .76)

We have recently conducted experiments in 5th grade classes (see Raatz, this volume). Once again we took the oportunity to calculate a regression equation. This time it takes the form:

Y' = 182 - 142 TTR - 0.83 XLS(W)
(Multiple R =.69)

Now these two equations differ, but they have to differ, since in the first case we are predicting 3rd grade difficulty, in the second we are predicting 5th grade difficulty.

The obvious way to check on the accuracy of our predictive formulae would be to administer a number of new texts to 3rd and 5th grade groups. We have not been inactive in the two years since we last presented a report here in Colchester, but there are still only two of us, and our enthusiasm for the C-Test, while enormous, is not unlimited. So we have not been able to do this yet. However we do possess a certain amount of evidence that the formulae work.

The first proof comes from the predictors themselves. Obviously they differ since we are predicting to a different grade level in each case. However both formulae use the same two indicators: the ones which correlate highest with the difficulty indices for the group involved. The relative importance of the indices in the prediction can be ascertained by examining the beta weights. These are virtually identical in both cases. TTR enters with .65 in the third grade and with .62 in the fifth grade. Mean sentence length in words has a beta weight of .39 in the third grade and .41 in the fifth. So not only are the same indicators being used, they also have the same importance in the two equations. Only eight texts were identical in the two investigations, the others were different for the two groups.

We can check the agreement between the two formulae by comparing the prediction of the text difficulty that they produce. Using the eight common texts we calculated difficulty predictions using both formulae and then calculated a rank correlation coefficient for agreement and came up with a value of .96. This is shown in Figure 1 which shows not only the very high agreement between the values, but also the fact that the only real difference between the two values consists of an additive constant, reflecting the fact that the texts become easier as the children grow older. In other words: these equations make identical predictions.

We therefore have three proofs for the accuracy of our predictions: the same two indicators enter into the regression equations as predictors; they have the same beta weights; and if the age increase is taken into consideration they produce the same conclu-

Figure 1 Difficulty prediction for L1 groups using regression equations

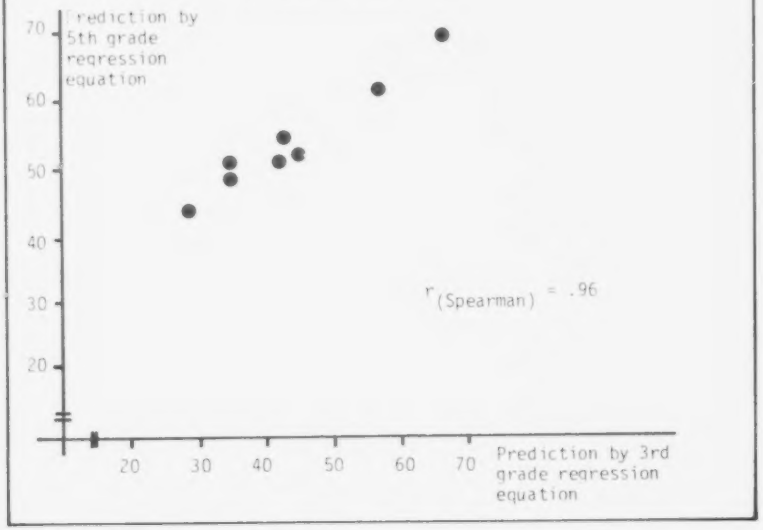
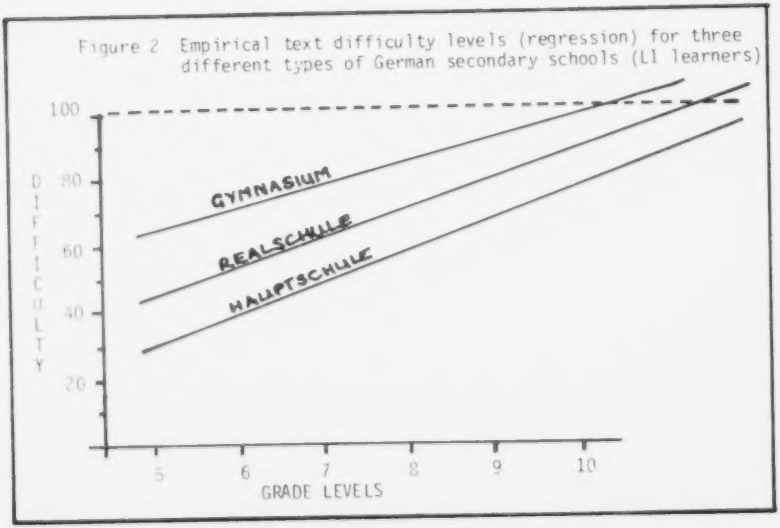


Figure 2 Empirical text difficulty levels (regression) for three different types of German secondary schools (L1 learners)



sions since they rank the texts virtually identically in terms of difficulty.

The fourth proof is empirical. For this C-Tests were administered to 19 classes in a Hauptschule, a Realschule and a Gymnasium. In the Hauptschule and the Realschule one class for each grade level between 6 and 10 was selected; in the Gymnasium in addition the 11th grade. In each of the 19 classes empirical difficulties were determined for 12 texts and for each text a prediction was calculated using the regression formula for the 5th grade. Finally the rank correlation between the empirically determined difficulty levels and the prediction produced by the regression equation was calculated.

The results are shown in Table 1.

Table 1

Results for the 12 texts in 19 school classes

School & class	N	Rank correlation between predicted and empirical difficulty values
H 6	17	.74
H 7	21	.54
H 8	19	.71
H 9	29	.72
H 10	20	.67
R 6	29	.83
R 7	23	.74
R 8	26	.70
R 9	27	.69
R 10	30	.82
G 6	26	.62
G 7	31	.57
G 8	34	.59
G 9	30	n.s.
G 10	27	n.s.
G 11	9	n.s.

All the rank correlations for the individual classes are significant with the exception of those for the Gymnasium classes 9, 10 and 11. In all three cases, however, the average total score is around 95%, so that there is no real variation in difficulty between the individual texts. The other correlations range between .54 to .83 with a median value of .70. When the empirical data are pooled the correlation between empirical text difficulty and predicted text difficulty rises to .85. Such positive confirmation of the predictive accuracy of our regression equation is rather surprising, since strictly speaking the regression equation developed for 5th grade

pupils is only applicable for this grade level. Yet it is obviously possible to predict the empirical difficulty ranking for texts in different kinds of schools and for different grade levels by using the regression equation derived initially for the 5th grade.

One reason for structuring this investigation in the way described was that we were interested in finding out whether C-Tests continued to decrease in difficulty as the pupils got older, and whether this decrease was linear. You may perhaps remember (cf. Raatz & Klein-Braley 1983) that we had shown such linear decreases in difficulty for the Hauptschule. Figure 2 shows the results of the investigation. For all three types of school the empirical facility values for the C-Tests become successively higher as the age of the pupils increases, and in all three cases the increase is linear.

This has very interesting implications for our regression equations. It is obviously possible to predict the RANK difficulties of texts used in C-Tests with a high degree of accuracy for older pupils even when the regression equation for the 5th grade is used. We have demonstrated that this technique works up to the 10th grade. However, since the regression equation for the 5th grade and that for the 3rd grade make virtually identical predictions, the 5th grade equation can obviously be used for all pupils between the 3rd and the 10th grade. Our data also show, however, that the regression lines for the C-Tests tend to hit the 100% mark at some point near the 11th grade. At the other end of the scale, it is not possible to use C-Tests, which are written tests, much earlier than the end of the 2nd grade in schools. This means that we have a regression equation which predicts the rank difficulty of texts for all relevant age groups of L1 German learners.

However we can go further than this. Two of the tests that were developed for the third grade L1 German learners were administered to our other two groups: to L2 learners and to LF learners. Süßmilch used test D3A in his pilot experiment with the children of foreign workers in Germany (L2 learners). Raatz chose to use test D3 with the LF learners at the Eurozentrum in Cologne.

The results of these tests show a very interesting phenomenon. As Table 2 shows, the rank difficulties of the tests were virtually identical to the rank difficulties derived for the L1 German learners. A further experiment showed that this applied to a group of British Army officers in Mülheim learning German too.

Table 2

Rank difficulties of texts for various groups

	TEXTS	3rd grade L1	3rd grade L1	3rd grade L1	Adults LF (Cologne)	Adults LF (Army)	3rd grade L2 (Turks)
D3	01	1	1	1	1	1	1
	02	2	2	2	2	2	2
	03	3	3	3	3	3	3
	04	4	4	4	4	4	4
	05	5	5	5	5	5	5

		3rd grade L1	3rd grade L2 (Turks)	3rd grade L2 (Greeks)	4th grade L2 (Turks)	3rd grade L2 (mixed)
D3A	01	1	1	1	1	1
	03	2	2	2	2	2
	20	3	3	3	3	4
	04	4	4	4	4	3
	17	6	6	5	5	5
	05	5	5			

In other words, the regression equation derived for the 5th grade L1 learners may also be used to determine the relative difficulty levels for both L2 and LF groups.

I should like to emphasise the word relative. The absolute difficulty levels can not be derived from our regression equation. However if one or two texts are administered to the relevant groups, the empirically determined difficulties for these texts can be used in conjunction with the regression equation to determine which texts are relatively easier, relatively more difficult, or relatively equal in difficulty with regard to the texts actually examined. This has the added advantage that the texts investigated empirically need not necessarily be included in the test. It will nevertheless be possible to develop tests for specific target groups which will have approximately the right level of difficulty for these groups. While, as I already stated, C-Tests have been shown to function adequately even if they are far too difficult or far too easy, there is no doubt that the best test is one with a medium level of difficulty for the subjects involved. Otherwise factors such as frustration or euphoria play far too large a role.

I think I am probably right in claiming that here for the first time language testers have a technique for estimating test difficulty in advance for the three major groups of test subjects, at least in one language, namely German. I would, however, make one reservation.

Our LF learners were an extremely heterogenous group. They came from a large number of different countries and spoke a large number of mother tongues. In addition their previous learning histories were unknown. The typical group of LF learners, however, is not like this. In the classroom we have a homogeneousmonolingual group. In addition this group learns the foreign language - normally at least - through a text book, or teaching materials. Such materials are structured and ordered so as to enable to learner to progress though various stages of vocabulary learning, syntax, grammar, communication in a pre-planned sequence. Whether our regression equation will work for groups in a controlled setting has not yet been examined. With L2 groups it seems not unreasonable to assume that they might have met the various components or elements of the language according to their normal frequency of occurrence in a similar fashion to the L1 learners. This assumption cannot be made for learners in a controlled, text-book-determined-language-learning environment. For this reason it may well be the case that the regression equation which is valid for all the other groups might fail to predict difficulty adequately here. The results for the British Army officers seem to contradict this assumption, butthese were very advanced LF learners, and at this stage the text book is no longer in control of the language learning process - the advanced learner is generally concerned with authentic language material.

It seems therefore that we have reached our aim in German. What about English?

Once again the largest group of subjects available to us is a German group: in this case LF learners, and most of our research has been concerned with our advanced students of English in Duisburg. This is, incidentally, the group for whom I personally really need techniques for advance determination of difficulty, since my own judgements of texts involved almost invariably turn out to be wrong. I nearly always overestimate the difficulty, and the test turns out to be too easy in the empirical trials.

Once again we follwed the same procedure: empirical test trials, calculation of the various indices, determination of the regression equation. We decided to use the C-Test values for the first group for whom the test had turned out to have an appropriate level of difficulty. These were the 67 students who were tested on entering the university in October 1982.

The regression equation for English texts for LF learners, calculated on the basis of 6 texts, once again uses the type-token ratio, but the syntax predictor here is the average sentence length in syllables. The equation takes the form:

$$Y' = 192 - 133\ TTR - 0.7\ XSL(S)$$

Table 3

Prediction for LF groups

Text	P'values	P values DU 82	P values DU 83	P values SF
GB07	97	94		88
GB11	66	66	75	55
GB17	80	83		67
GB20	60	58	62	50
GB19	63	59	64	44
GB18	56	56		55
GB30	81		81	

$r_{12} = .987$ $r_{13} = 1.00$ $r_{14} = .930$

Here the beta weights are -.65 for the TTR and -.83 for the mean sentence length in syllables. The multiple correlation between these two indices and the empirically determined text difficulties is a staggering .98!

Again, however, we need to determine the accuracy of the predictions obtained. One proof lies in the data obtained by using some of the same texts in the tests run for a new group on October 4th 1983. For the text GB 20 the predicted value is 60 and the empirically measured value is 62. For text GB 19 the predicted value is 63 and the empirical value 64. For text GB 11 the predicted value is 66 and the empirical value 75. The fourth text in this test (GB 30) had never been used before, but had a predicted value of 81. Its empirical value turned out to be precisely 81 (cf. Table 3). The rank correlation coefficient between the predicted difficulty levels and the empirically determined levels is 1.00.

This seems to indicate that the regression equation calculated for incoming students in Duisburg works very well for this group. I was fortunately able to make a second check. Around three weeks ago I received a set of tests from Hilkka von Essen which she had admin-

Figure 3 Predicted and empirical difficulty values for Finnish LF group

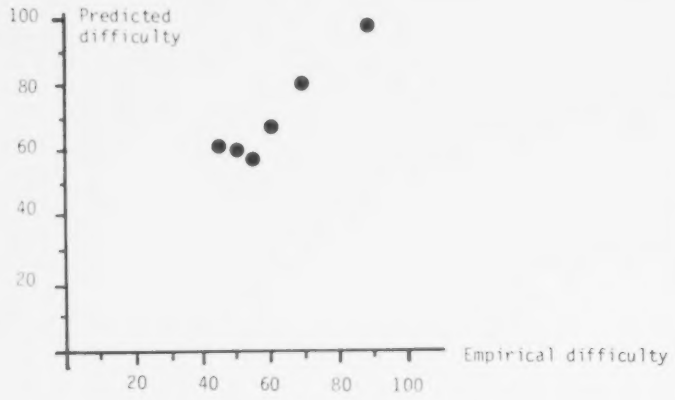
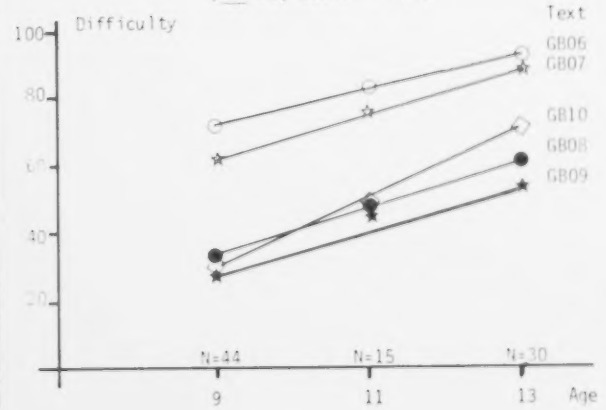


Figure 4 Linearity of difficulty for English L1 groups (not regression lines)



istered to her students in Finland (SF group in Table 3). This was the same test which had originally been used to develop the regression equation. Here therefore we can check on the accuracy of the prediction for a different LF advanced group. The correlation between the empirical values and the predicted values in this case is .93, and as Figure 3 shows, again we have a case where obviously what is needed is simply an additive constant.

A final confirmation is given by the use of the regression equation for LF groups to predict text difficulty for L1 groups. Here three factors have changed: the group for whom the prediction is to be used is L1 and not LF; the age levels are different (9 to 13 year olds and not 20 year olds) and the texts are entirely different. Again, it is possible to show that the decrease in difficulty is linear (Figure 4). In all cases the correlation between the predicted difficulties and the actual difficulties is between .6 and .7. With only 5 texts involved this value is, of course, not statistically significant. But as examination of the data shows, (Figure 5), the prediction obtained from the regression equation successfully divide the texts into two groups, and they do this for all three age levels involved. It would therefore be possible to work with this equation until we have more stable data from larger L1 groups.

The question I have started to ask myself as a result of these findings for both English and German is whether the basic equation ought perhaps to be the one for the L1 groups. But more research will be necessary before this question can be answered.

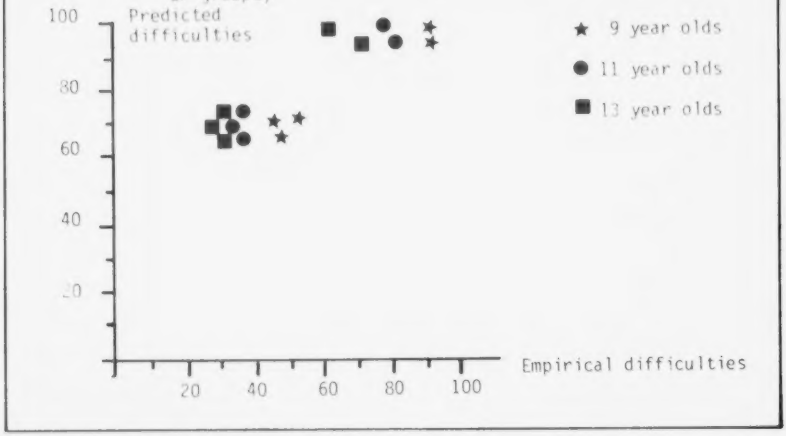
What conclusions can be drawn from all this? The first is an eminently practical one, namely that one of the major problems with the C-Test, that of advance prediction of text difficulty, is well on the way to being solved. In our future work we shall be making use of these equations, and at a future IUS meeting we shall report on their performance.

The second conclusion is more theoretical. It seems to me that these results imply that there is such a thing as absolute text difficulty, which I propose to call C-processing difficulty. We have shown both for English and German that the C-processing difficulty decreases in a linear fashion as the L1 subjects get older. We have shown that regression equations using text statistics derived for specific subgroups can be used with a fair degree of predictive accuracy over wide age ranges. We have also shown that the same regression equations can be used for L1 and LF groups, and that for German they can also be used for L2 groups.

One of the things that fascinate me about our results is the parallelism for the different learner groups. This seems to me to have all sorts of implications for language acquisition research. However I have, as yet, only begun to scratch the surface of this problem. Instead I should like to finish by reporting our newest idea.

You will remember that the advantages of the Rasch model were

Figure 5 Empirical and predicted difficulty levels for 9, 11 & 13 year olds (using equation derived for LF groups)



firstly that the latent trait of the subject and the difficulty of the items can be measured unidimensionally on an interval scale and secondly that the latent trait can be measured independently of the items and vice versa. In the Rasch model the parameters ξ and ε are related to each other in a complex multiplicative function (logistic function). For quite a while now we have been searching for a way to use the Rasch model for our data. The observed score for any individual on a C-Test must be dependent on the one hand on the C-processing difficulty of the text, and on the other on the individual's position on the language learning continuum. A simpler formal model than the Rasch model for this relationship could be an additive one:

observed score = subject ability + text level + error.

In October 1983 at a psychology congress in Düsseldorf Moosbruger and Müller presented their CLA model, which has precisely this basic equation. The model is analogous to the Rasch model; it has exactly the same qualities and is intended to be used for at least ordinally scaled data.

Here the parameters are $X_{ij} = \xi_i + \lambda_j + \varepsilon_{ij}$ where ξ_i is subject ability (latent trait) and λ_j is text difficulty level (C-processing difficulty).

We intend to use this model and hope that it will enable us to disentangle C-processing difficulty and subject ability, positing a latent trait which we might call general language competence. If this is possible, then the C-Test which began its life simply as a toy for two incurable test freaks may well turn out to represent a significant advance in research into language learning.

Notes

I should like to thank Werner Debertin who administered the tests as part of his thesis for the First State Examination for Teachers.

Bibliography

Alderson, J. C. 1979."The cloze procedure and proficiency in English as a foreign language". TESOL Quarterly 13,2: 219-226.

Bartenwerfer, H. and U. Raatz. 1979. Einführung in die Psychologie 6. Methoden der Psychologie. Wiesbaden: Akademische Verlagsgesellschaft.

Bensoussan, M. 1982. "Testing the test of advanced EFL reading comprehension. To what extent does the difficulty of a multiple-choice reading comprehension test reflect the difficulty of the text?" SYSTEM, 10: 285-290.

- 1983. Changing the difficulty level of multiple-choice reading comprehension questions. Paper presented at the ACROLT Symposium, Quiriyat Anavim, Israel.

Herdan, G. 1966. The advanced theory of language as choice and chance. Berlin: Springer.

Klein-Braley, C. 1981a. Empirical investigations of cloze tests. Dissertation, Duisburg.

- 1981b. The determination of text difficulty. Paper presented at the 12th Jahrestagung of the Gesellschaft für Angewandte Linguistik, Mainz.

Moosbrugger, H. 1983. Neuere psychologische Testtheorie und ihre Anwendung in der Psychodiagnostik. Paper presented to the 12th Kongreß für Angewandte Psychologie des BDP, Düsseldorf.

Oller, J.W.Jr. 1979. Language tests at school. London: Longman.

Pugh, A.K. 1981. "Text analysis, comprehension and reading behaviour". In: Hopkins, E. and R. Grotjahn. (eds). 1981. Studies in language teaching and language acquisition. Quantitative linguistics 9. Bochum: Brockmeyer. 109-128.

Stokes, A. 1980. The reliability of readability formulae. In: Pugh, A.K., Lee, V.J. and J. Swann. (eds). 1980. Language and language use. London: Heinemann. 345-357.

A REGRESSION ANALYSIS OF DIRECT AND INDIRECT MEASURES OF ENGLISH AS A SECOND LANGUAGE WRITING COMPOSITIONS

Kyle Perkins (Southern Illinois University)

Abstract

This study attempted to establish the strength of relationships between direct and indirect measures of attained English as a Second Language writing proficiency. The Test of Ability to Subordinate (TAS) was the indirect measure and holistic scores for free writing samples the direct measure. Substantial magnitudes of correlation were obtained between the TAS and the holistic evaluations. A stepwise multiple regression analysis indicated that both direct and indirect measures of writing accounted for substantial amounts of variance in a direct measure variable. Explanations are offered to account for the remaining variance in the dependent variable. Limitations of the study are discussed, and suggestions are offered for future research using this paradigm.

Introduction

There is considerable debate in the composition research community concerning the efficacy of direct and indirect tests of attained writing proficiency. Some researchers, like Michael and Shaeffer (1979) and Diederich (1974), advocate the use of multiple-choice tests of writing.

Michael and Shaeffer stated:

> A short multiple choice examination requiring knowledge of principles about grammar, usage, choice of words and phrases, and syntax such as the TSWE (Test of Standard Written English) or a subtest of the CSUC-EPT (California State University and Colleges English Placement Test) reflecting similar competencies provides a more accurate prediction of success in a basic English composition course emphasizing writing skills than does an essay examination requiring the actual demonstration of writing skills (p 132).

In a similar vein, Diederich noted:

> My final word of advice on such tests (objective tests) is not to despise them. Objective test items can easily, quickly and reliably test a student's knowledge of the rules and conventions of English. In writing, if a student is not sure that he knows how to use a certain construction he can change his sentence to avoid using it, but in an objective test, you can

> give him a sentence with that construction in it, and he has to decide whether it is correct or incorrect. Arguments over whether answering objective items 'is the same thing as' actual writing or speaking are futile. For that matter, writing one essay is not 'the same thing as' writing another essay, even on the same day; we have seen that even the most carefully determined grades on such essays rarely correlate higher than 0.70. Then the astonishing thing is that scores on a good objective test of English usage often correlate about 0.70 with averages of the two essay grades. It does not matter that they do not 'really' measure 'the same thing.' If students who are good at one tend to be good at the other, and vice versa, then it is a good indicator of proficiency in written English. Call it an editing test if you like, but I can promise you that students who do well on it tend to be good writers (pp 79-80).

But other researchers like Braddock, Lloyd-Jones and Schoer (1963) and Breland and Gaynor (1979) have criticized objective tests for their lack of validity and the impressions about the low valuation of writing which they convey.

Braddock, Lloyd-Jones and Schoer:

> The most serious charge against multiple-choice tests is their lack of validity. Not only do they require the examinee to perform the behavior being measured - he does no actual writing; but these tests also make no attempt to measure the 'large elements' of composition, even indirectly (p 42).

Breland and Gaynor:

> Indirect measures lack face validity and credibility among members of the English profession and educators generally, and they tend to deliver a message to students that writing is not important (p 127).

For further discussion on the correlation between direct and indirect test scores in composition, see Davis, Scriven and Thomas (1981); Hogan and Mishler (1980); Moss, Cole and Khampalikit (1982); and Stiggens (1982).

Research Questions

The research reported in this paper concerns (1) the strength of relationship between (a) direct and indirect measures of attained English as a second language (ESL) writing proficiency; (b) years of ESL study and direct and indirect measures of writing proficiency and (2) the amount of variance which a direct and an indirect measure and years of ESL study will account for in a direct measure of ESL writing.

More specifically, the following research questions were addressed:

1 Would a direct and an indirect measure of attained ESL writing proficiency correlate significantly?

2 Would years of ESL study and direct and indirect measures of ESL writing correlate significantly?

3 How much variance would years of ESL study, a direct and an indirect measure account for in a direct measure of ESL writing?

Procedures

Subjects This research was conducted in the Department of Linguistics, Southern Illinois University. The subjects were sixty seven undergraduate foreign language students enrolled in Basic Composition for Foreign Students. Native languages represented in the sample included Lao, Arabic, Farsi, Malay, Korean, Vietnamese, Thai, Chinese, Japanese, Igbo, Greek, Portuguese, Spanish, Norwegian and Gujerati.

Materials and Procedures At the beginning of a fifteen-week semester the Test of Ability to subordinate (TAS) was administered to the subject pool. During the same week the subjects wrote an expository composition which was graded holistically. During the testing procedure certain demographic data were collected, including the number of years each subject had studied English. During the final week of class the TAS was administered again and another expository composition was collected and graded holistically.

The TAS originates from a school of thought that holds that enhancement of syntactic fluency will also be reflected in improvement in overall writing ability. The TAS is considered, for this research, an indirect measure of writing, since it entails no text construction or actual composing. The following description of the TAS is found in the Teacher's Manual:

> The Test of Ability to Subordinate was designed primarily as a diagnostic instrument ot measure the ability of non-native speakers of English to control certain structures of subordination in writing. It was developed to fill the need for an easily scoreable, objective test that would measure students' ability to manipulate particular grammatical structures generally regarded as indicative of mature writing . . . the primary diagnostic purpose of the TAS is to identify which of nine structures of subordination students can manipulate and which they cannot, but a careful analysis of responses will highlight difficulties, if

any, with subject-verb agreement, plurals, tenses and articles. A teacher can use the information from the TAS to group students in terms of their strengths and weaknesses in subordination and to develop a writing program that will enable them to use these structures effectively as well as deal with other grammatical problems.

The essays were graded holistically by the author and a colleague. A five-point scale was employed: 5:excellent - 1:bad. Each paper was evaluated for form and content. To increase the variance the author summed all scores (rather than average the two evaluators' marks) for each paper. For example, say paper #33 received the following scores: Grader A, form 3; content 3; Grader B, form 4; content 5. The holistic score recorded for paper #33 was 15. The scoring was done independently and anonymously by each reader in an attempt to preclude any halo effect or contamination from affecting the results.

Analyses

The summary statistics and estimates of reliability for the writing measures are reported in Table 1. Both the TAS and holistic scoring exhibited acceptable coefficients of reliability for research of this nature.

TABLE 1

SUMMARY STATISTICS

Variable	Mean	S.D.	Type of reliability	Coefficient of reliability
L2 years	8.67	4.41		
Pre-TAS	31.85	7.41	KR-21	.85
Post-TAS	35.21	6.64	KR-21	.85
Pre-holistic	11.28	3.61	Interrater	.90
Post-holistic	11.25	3.58	Interrater	.92

The data generated for each variable were submitted to the Pearson product-moment correlation procedure. The correlation matrix is shown in Table 2.

TABLE 2

CORRELATION MATRIX

r_{xy}

	1	2	3	4	5
1. L2 years	1.00				
2. Pre-TAS	.48*	1.00			
3. Post-TAS	.32*	.82*	1.00		
4. Pre-holistic	.27	.75*	.69*	1.00	
5. Post-holistic	.19	.56*	.51*	.55*	1.00

* $p < .01$, one-tailed test

Since the magnitude of any correlation coefficient is affected by the reliabilities of the measures of the X and Y variables, it is necessary to correct the correlation coefficients to hold reliability constant - the so-called correction for attenuation procedure. The disattenuated correlation matrix is presented in Table 3.

It can be seen in Table 3 that all correlations are significant at the .01 level but for L2 years, pre-holistic, and post-holistic. However, with large sample sizes such as this one (df=65) it is possible for comparatively small correlation coefficients to be found significant. Since coefficients above 0.50 are generally thought to be meaningful for research purposes, the correlations of 0.48 between L2 years and pre-TAS, and 0.32 between L2 years and post-TAS will be ignored, although they are significant at df65.

TABLE 3

CORRELATION MATRIX CORRECTED FOR ATTENUATION

r_{CA}

	1	2	3	4	5
1. L2 years	1.00				
2. Pre-TAS	.48*	1.00			
3. Post-TAS	.32*	.96*	1.00		
4. Pre-holistic	.27	.86*	.79*	1.00	
5. Post-holistic	.19	.64*	.58*	.60*	1.00

* $p < .01$, one-tailed test

A stepwise multiple regression analysis was conducted to determine how much variance the pre-TAS, post-TAS, pre-holistic and L2 years variables would account for in the post-holistic variable. The post-holistic scores were chosen as the criterion variable for the following reasons: (1) a direct writing sample at the end of a fifteen week semester of composition should be indicative of student writing proficiency and (2) like Braddock, Lloyd-Jones and Schoer, and Breland and Gaynor, the author prefers direct instead of indirect measures of writing. For ESL placement and proficiency testing the author prefers holistic evaluation of direct writing samples for the following reasons.

Holistic scoring has the highest construct validity of any composition evaluation scheme, when overall attained writing proficiency is the construct to be assessed. With holistic grading (1) the entire text is evaluated; (2) the grader can determine whether a thesis has been clearly stated, developed and supported or whether an issue has been clearly raised and sufficiently resolved; whether sufficient support and development (i.e., redundancy in the communication science sense) have been provided for the reader; whether the writer has tried to

accommodate the reader in terms of how much (how little) the writer thinks the reader knows about the subject (i.e., has the writer tried to accommodate the intended audience?); whether there is grammatical and lexical cohesion and overall coherence; whether one paragraph is 'hinged' smoothly to preceding and succeeding paragraphs through the accomplished use of transitional words, phrases and structure vocabulary; whether new information has been converted into given information that forms the base for predicating additional new information; and finally whether the piece 'makes sense'.

The results of the stepwise multiple regression are presented in Table 4.

TABLE 4

STEPWISE MULTIPLE REGRESSION

Criterion Variable: Post-holistic

Step	Source	R	R^2	R^2 INC	R^2 ADJ	DF	F INC	F TOT
1	Pre-TAS	.56	.314	.314	.303	1,65	29.75	29.75
2	Pre-holistic	.59	.352	.038	.332	1,64	3.75	17.38
3	LZ years	.60	.356	.004	.325	1,63	.39	11.61
4	Post-TAS	.60	.358	.002	.317	1,62	.19	8.64

Results and Discussion

As Table 3 indicates, the highest correlations obtained were 0.96 between pre- and post-TAS; 0.86 between pre-TAS and pre-holistic; 0.79 between post-TAS and pre-holistic; and 0.64 between pre-TAS and post-holistic. The largest coefficient represents a 92% overlap the two administrations of the TAS and could be interpreted as a coefficient of test-reliability. The three remaining coefficients which were cited could be viewed as

evidence that an indirect measure of writing, TAS, correlates substantially with a direct measure of writing, the holistic evaluations. The coefficient of 0.86 between pre-TAS and pre-holistic is correlational evidence of the TAS's concurrent validity with a 74% variance overlap. Pre-holistic and post-holistic correlated 0.79 - the 62% overlap being another correlational piece of evidence that a direct and indirect measure of writing can share substantial amounts of variance. And finally, there is a less substantial correlation of 0.64 which obtained between pre-TAS and post-holistic - a measure of the TAS's predictive validity. In sum, there seems to be evidence from the disattenuated correlation matrix that a direct and an indirect measure of ESL writing proficiency will share substantial magnitudes of covariance.

A more rigorous examination can be found in the stepwise regression analysis. The purpose of a multiple regression is to analyse the relationship between a dependent criterion variable (in this study, the post-holistic scores) and a set of independent, predictor variables (in this study, L2 years, pre-TAS, post-TAS and pre-holistic). For purposes of this research the linear dependence of post-holistic on L2 years, pre-TAS, post-TAS and pre-holistic is summarized and decomposed.

Table 4 illustrates the order in which the predictor variables entered the equation and the amount of variance accounted for by itself and cumulatively in linear combination with other predictor variables. Pre-TAS was the most robust predictor accounting for 31.40% of variance in the post-holistic variable. Pre-holistic accounted for an additional 3.80% of variance, the two variables together accounting for only 0.60% of the dependent variance (R^2 INC .004 - .002 = .006 x 100 = 0.60%).

At first blush it appears that the stepwise multiple-regression analysis establishes predictive validity for the TAS and the holistic evaluation and fails to establish concurrent validity for the TAS. These interpretations can be drawn by examining the F-statistics; however, the disappointing aspect of Table 4 is that the four predictor variables account for only 35.80% of the variance in the dependent variable, leaving 64.20% of the dependent variance unaccounted for by the predictor variables. It would be facetious to expect the predictor variables to account for 100% of the reliable variance in the dependent variable. On the other hand, the researcher believes that if the TAS were a really robust predictor variable, as well as the pre-holistic, 64.20% of the variance would not have been left unaccounted for. The point being made here is that for this kind of research, the magnitude of the r, R, and R^2 coefficients is more important than their levels of tabled value significance. Perhaps the findings can be explicated in the following section of the paper.

Limitations of the Study

There are at least two major limitations to this study. First, only one essay was evaluated for the pre-holistic scores and only one essay for the post-holistic scores. At least two essays should have been evaluated for each data collection for the following reason. Research has shown student writers vary significantly in their day-to-day performance in the composing process (Anderson 1960; Arthur 1979; Braddock-Lloyd-Jones and Schoer 1963; Kincaid 1953). The elicitation of multiple writing samples from the subject pool at various times on varying topics would have controlled for the fact that writing ability may vary with topic and time of day. Secondly, the different essays should have been elicited from different modes of discourse, because research has shown that writing students' ability varies across discourse forms (Arena 1975; Britton et al 1975; Kincaid 1953; Kinneary 1971; Lloyd-Jones 1977). Had these limitations been attended to, there might have been a significant difference between the pre-and post-holistic means (Table 1) and substantially more variance accounted for in the dependent variable (Table 4).

Although a multiple R or 0.60 (Table 4) is not shamefully low, the author would prefer to have more variance in the dependent variable accounted for by the predictor variables. There are five potential explanations as to why the multiple R was not larger: 1) the dependent or independent variables were not reliable; 2) the predictors were not robust; 3) there was a trait present which was not tapped by the measures; 4) there was redundancy present in the independent variables; 5) there was multicollinearity present in the independent variables. The first two explanations can be rejected out of hand, because the lowest reliability estimate for any variable was 0.85% and two independent variables did exhibit significant F-statistics in the regression equation. It would be impossible to empirically test claims three and four without a replication of this design with different instrumentation. The remaining factor, multicollinearity, is offered then as an explanation for the variance which was unaccounted for in the dependent variable. Multicollinearity is present when some or all of the independent variables are very highly intercorrelated. An inspection of Table 3 reveals this condition to be the case for this particular study. When the independent variables are very highly intercorrelated, the reliability of the relative importance indicated by the partial regression coefficients is reduced.

Conclusion

This study attempted to establish the strength of relationships between direct and indirect measures of attained ESL writing proficiency. The TAS was used as an indirect measure of writing and holistic evaluations of form and content were used as direct measures. Both the direct and indirect measures exhibited high estimates of reliability. Substantial magnitudes

of correlation were obtained between the two administrations of the TAS and the TAS and the holistic evaluations. A stepwise multiple regression analysis indicated that the pre-TAS and pre-holistic scores accounted for a significant amount of variance in the criterion variable, but all the predictor variables cumulatively accounted for only 35.80% of the variance in the dependent variable. Multicollinearity was suggested as the factor that accounts for lack of a larger multiple R. It was suggested that the lack of multiple writing samples from different discourse modes may pose limitations to the external validity of the study.

Such a study needs to be replicated to determine whether the results reported herein are valid. Other indirect measures such as the TSWE, the Cooperative English Test Effectiveness and Mechanics subtest, or the Revision and Editing Test (Matalene 1982) could be used as independent variables and holistic scores or primary trait scores, the criterion variables.

References

Anderson, C.C. 1960. The New STEP Essay Test as a Measure of Composition Ability. Educational and Psychological Measurement 20, pp 95-102

Arena, L. 1975. A Method for Improving the Writing Skills of Foreign Students in University-Level Expository English Composition Courses. In Burt, M.K. and Dulay, H.C. (eds) On TESOL '75: New Directions in Second Language Learning, Teaching, and Bilingual Education Washington D.C.: TESOL, pp 281-291

Arthur, B. 1975. Short-Term Changes in EFL Composition Skills. In Yorio, C.A., Perkins, K. and Schachter, J. (eds) On TESOL '79: The Learner in Focus Washington D.C.: TESOL, pp 330-342

Braddock, R., Lloyd-Jones, R. and Schoer, L. 1963. Research in Written Composition Urbana, IL: NCTE

Breland, H.M. and Gaynor, J.L. 1979. A Comparison of Direct and Indirect Assessments of Writing Skill. Journal of Educational Measurement 16, pp 119-128

Britton, J., Burgess, T., Martin, N., McLeod, A. and Rosen, H. 1975. The Development of Writing Abilities (11-18) London: Macmillan Education Ltd

Davis, B., Scriven, M. and Thomas, S. 1981. The Evaluation of Composition Instruction Pt. Reyes, CA: Edgepress

Diederich, P.B. 1974. Measuring Growth in English Urbana, IL: NCTE

Hogan, T. and Mishler, C. 1980. Relationships Between Essay Tests and Objective Tests of Language Skills for Elementary School Students. Journal of Educational Measurement 17, pp 219-227

Kincaid, G.L. 1953. Some Factors Affecting Variations in the Quality of Students' Writing Unpublished doctoral dissertation, Michigan State University

Kinneavy, J.L. 1971. A Theory of Discourse Englewood Cliffs, NJ: Prentice-Hall

Lloyd-Jones, R. 1977. Primary Trait Scoring. In Cooper, C.R. and Odell, L. (eds.) Evaluating Writing: Describing, Measuring, Judging Urbana, IL: NCTE, pp 33-66

Matalene, C. 1982. Objective Testing: Politics, Problems, Possibilities. College English 44, pp 368-381

Michael, W. and Schaeffer, P. 1979. A Comparison of the Validity of the Test of Standard Written English (TSWE) and of the California State University and Colleges English Placement Test (CSUC-EPT) in the Prediction of Grades in a Basic English Composition Course and of Overall Freshman-Year Grade Point Average. Educational and Psychological Measurement 39

Moss, P., Cole, N. and Khampalikit, C. 1982. A Comparison of Procedures to Assess Written Language Skills in Grade 10, 7, and 4. Journal of Educational Measurement 19, pp 37-48

Stiggins, R. 1982. An Analysis of Direct and Indirect Methods of Writing Assessment. Research in the Teaching of English 16, pp 57-69

THE FACTORIAL VALIDITY OF C-TESTS

Ulrich Raatz (Colchester, October 1983)

1 Introduction

It was in Colchester at the Fourth International Language Testing Symposium of the IUS, almost exactly two years ago, that Kris Klein-Braley and I first presented the C-Principle and the earliest investigations with C-tests in English and German to an expert audience. Since then the interest in this type of test among language testers, psychologists and educationists has been steadily increasing. C-Tests have been used, or are being used, in Germany, England, Belgium and Finland. Rebecca Valette has spread the news in the United States of America. Kris was invited to Israel to talk to the Israeli Committee on Language Testing about C-Tests. Interested colleagues have taken up our challenge and developed C-Tests for their own target groups, for instance Edgar Süßmilch for the children of immigrant workers in Germany and Rüdiger Grotjahn for university students of French. First test forms have been developed for Spanish, Greek and Turkish. In Cologne at the Eurozentrum, where students from all over the world come to learn German, a C-Test is part of the regular test battery used for placement of incoming students, together with traditional discrete-point item tests.

In the last two years, Kris Klein-Braley and I have presented papers on the C-Test at around 12 national and international congresses. At this point of time we have conducted a total of 21 different empirical investigations into C-Tests and published the results. Other investigations are currently in progress. Yet others have been completed but have not yet been published, and it is the result of one of these investigations that I want to present in this paper.

In our research into C-Tests over the past two years we have been concerned with three main questions:

1. the examination of the reliability and validity of C-Tests for a variety of different groups;
2. investigations into the suitability of C-Tests for classification decisions;
3. examination of the linguistic and psychometric theory underlying C-Tests.

We have administered German and English tests to native speakers of these languages, to learners of English and German as a second language, to foreign language learners. Our subjects range from pupils in the third to tenth grades, and from the sixth form level to adults. In all, around 3 000 subjects have taken tests. All this research activity has not been able to discredit the C-Principle and the C-Tests.

2 Previous validation of C-Tests

In all our investigations into C-Tests up to now one of the most important questions has been that of validity. What do C-Tests measure?

Since C-Tests use several unsystematically selected texts, whose subject matter is as varied as possible, since in each case the mutilated words are a representative sample of all the words in the texts, and since, in addition, they are based on a linguistic theory, namely Oller's construct of pragmatic expectancy grammar, we claim that C-Tests are content valid tests.

We have used two different criteria to determine empirical validity. In 12 investigations into L1 groups we used teacher ratings as our criterion. This was either the last school grade given on a regular report or an estimation of current performance. Sometimes the global performance was evaluated, sometimes individual components such as orthography or grammar. In such cases the individual scores were averaged to obtain the criterion. Table 1 shows the distribution of the 12 validity coefficients.

Table 1

Distribution of validity coefficients for previously published investigations (Raatz and Klein-Braley 1983)

r_{tc}	f
0 - .10	1
.11 - .20	
.21 - .30	
.31 - .40	
.41 - .50	2
.51 - .60	2
.61 - .70	1
.71 - .80	4
.81 - .90	2
.91 -1.00	

In 11 investigations with LF groups we were able to make use of scores obtained from other test batteries as criteria. These tests were being used in placement procedures. Table 2 shows the results for these investigations.

Table 2

Results of validity studies for LF groups
(Raatz and Klein-Braley 1983; Grotjahn 1983)

	Duisburg	Köln	Bochum
Language	English	German	French
Population	German students of English	Foreign students	German students of French
Criterion	DELTA-Test 6 Subtests	Euro-Test 4 Subtests	Placement Test F 7 subtests
N of investigations	8	2	1
Validity	.77 - .91	.82/.85	.87

The results for the French C-Test are new. They were reported at the Congress of the German Association of Applied Linguistics by Rüdiger Grotjahn.

Factorial validity has been determined so far in 5 investigations. In the two studies performed at the Eurozentrum in Cologne a factor analysis was calculated over the C-Test and the four subtests of the Eurotest. In both cases a general factor emerged which explained around 70 % of the variance, and the C-Tests had their highest loadings on this factor. Similar results emerged in three investigations in Duisburg for students of English where factor analyses were performed over the C-Tests and the six subtests of the DELTA Test. However, since the number of variables involved in each case was relatively small, these results cannot be regarded as definitive.

We can sum up the results of these investigations by drawing the following conclusions:

1. In the classroom setting the C-Tests produce the same judgments as the experienced teacher in L1 teaching.

2. The C-Tests measure the same thing that complex language test batteries measure in LF-teaching.

However, the question of what the C-Test actually measures has still not been answered. The pilot study to be reported here is a first attempt to provide an answer to this question.

3 The purposes of this investigation

In this study we investigated the convergent, divergent and factorial validity of a German C-Test in two parallel versions for German pupils in the 5th grade. The individual questions we hoped to answer were as follows:

1. What is the correlation between the C-Tests and the school grades in German, grammar and English and with other language tests? (convergent validity).

2. What is the correlation between the C-Tests and intelligence and concentration tests? (divergent validity).

3. What factor structure is produced by the tests used, and what is the implication for the factorial validity of C-Tests?

In view of comments I have previously made (cf. Raatz 1981) about the use of factor analysis in attempting to prove theories of language I should like to point out explicitly that in this study factor analysis is used as a descriptive technique. No hypotheses about an expected factor structure had been set up in advance.

4 The sample

In Mülheim, a smaller town near Duisburg, a school centre has been established where the three traditional types of secondary schools in Germany share a communal campus. These school types are Hauptschulen, Realschulen and Gymnasien. Roughly speaking the Hauptschule is equivalent to the secondary modern school, the Gymnasium to the grammar school, and the Realschule lies somewhere in the middle. It is, however, much more common than the secondary technical school in England.

Our investigation was performed in one 5th grade form from each type of school. The results were then pooled for the analysis. We hoped (by pooling) to avoid the restriction of variance which would necessarily occur if we studied only pupils from one of the three types of secondary school, as compared to the variance in the 4th grade form, the last form of the primary school, where the pupils are still together.

The structure of the sample is shown in Table 3.

Table 3

The structure of the sample

	N	N boys	N girls	mean age	Number of non-German children
Hauptschule	18	7	11	11	4
Realschule	24	13	11	10,9	1
Gymnasium	33	20	13	10,8	1
all	75	40	35	10,9	6

A total sample size of N = 75 is rather a small group for a factor analysis. For this reason we also performed a cross-validation study, by examining the results obtained for the two halves of the sample separately. Even so the results of the study are more indicative than conclusive and should be interpreted with the necessary caution.

5 Variables

The variables listed in Table 4 were used in the study.

Table 4

List of the variables included in the study

1. German C-Test in 2 parallel versions A and B
2. School grade in German
3. Teacher rating in grammar
4. School grade in English
5. School grade in mathematics
6. Diagnostic German Test (DTD 4-6) with the subtests
   (PW) PV Passive Vocabulary
   (AF) FA Finding Analogies
   (TS) TS Text Structuring
   (IV) IC Instruction Comprehension
   (LV) RC Reading Comprehension
   (AW) AV Active Vocabulary
   (in brackets the German abbreviations)
7. Orthography test DRT 4-5
8. Non-verbal intelligence test FDA 3-6
9. Concentration test d2

The two parallel forms of the C-Test (A and B) contain 6 texts each, and each text has 20 mutilated words. Eight texts had already been tried out in test forms, 4 texts were new. Administration of the C-Test took 40 minutes.

The school grades for German, English and mathematics were those given at the end of the last half-year. In Germany school grades can, in principle, vary from 1 (very good) to 6 (beyond the pale), and they are normed on the performance of the individual class. In the calculations the poles were reversed in order to avoid negative correlations or factor loadings.

Grades for grammar are not envisaged in the normal school grading scheme. However, here too the teachers used the familiar school grades as rating categories.

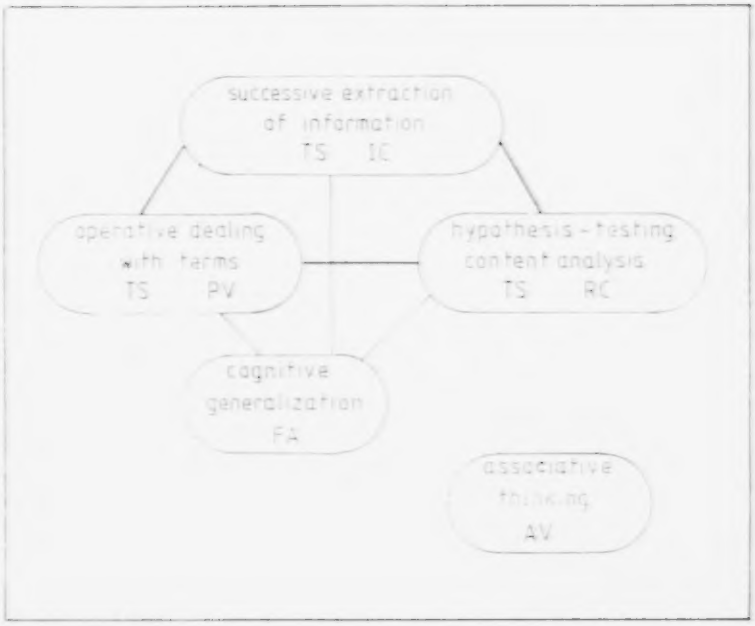
The Diagnostic German Test (DTD 4-6) was developed in 1980 by NAUCK and OTTE. This test takes as its model the assumption that all achievements in German mother tongue teaching can be attributed to 6 language-bound fundamentals, which are operationalised in the 6 subtests of the DTD listed and described below.

1. PV Passive vocabulary. Here the pupil is expected to find familiar words which are distorted in some way. To do this he must use his passive vocabulary. (22 items, 8 minutes.)
2. FA Finding analogies. Here comparisons, classifications and relations are given and analogies must be discovered. (22 items, 8 minutes.)
3. TS Text structuring. Simple sentences whose words are printed without spacing between them must be structured by inserting vertical lines. (22 items, 10 minutes)
4. IC Instruction comprehension. Here instructions of the type often found in instruction sheets or school workbooks have to be followed. The difficulty in processing lies in the connection of simple actions by complex operators. (17 items, 12 minutes.)
5. RC Reading comprehension. Two texts are given (mini-stories) with a number of 4-choice multiple-choice questions. Choice of the correct, or the best response is determined by the ability to understand the text without help. (18 items, 16 minutes.)
6. AV Active vocabulary. The pupil has to find as many words as he can which begin with the letter given. This indicator for word fluency is primarily dependent on the range of vocabulary available to the individual since the time limits are generous. (2 attempts, 2 minutes each.)

The authors combine one or two of these fundamentals to 5 part constructs, which form a model structure of language learning, as shown in Fig. 1.

Figure 1

The structure of part constructs in language learning
(Nauck and OTTE, 1968)



The three constructs at the top, which all contain the fundamental TS (text structuring), have the greatest importance in this structure. The OTO 4-6 test demands the ability to perform pragmatic operations and analyses on language; it does not demand the ability to describe the functions and working of language in a cognitive fashion. For this test two parallel forms are available. The total score for the test and the result for each individual subtest were entered into the analysis as a raw score.

The orthography test ORT 4-5 was developed by Meis (1969). The test uses a format which we in Germany call the Lückendiktat. The subject receives the text in a form similar to the rational deletion cloze test. However in this type of test the text is read to him and he inserts the critical words into the blanks. The score is the sum of the correctly written words, and was entered into the analysis as a raw score.

In order to measure intelligence I selected the non-verbal intelligence test FDA 3-6, which is also available in two parallel forms (Wendeler 1973). This test measures chiefly the general intelligence factor, together with the space factor. It has 3 subtests with 21,21 and 20 items respectively and each subtest has an time-allowance of 15 minutes.

In the first subtest a sequence must be continued using one of the 4 alternatives given. In the second subtest the figure which does not fit must be discovered. In the third subtest the missing figure in a matrix must be restored. Here again the raw score was used for analysis.

The concentration test d2 (also called the attention-stress test) was developed by Brickenkamp (1972). The subjects receive a test sheet with 14 lines of text; each line has 47 characters (d, p). The subject must cross out every d which has two dashes with it. He has 20 seconds for each line. The final score is the number of characters attempted minus the number of errors.

6 Test administration

The test sessions took place in a three week period during April 1983. For each group several sessions were necessary; in each case the first lesson periods were used.

7 Results

7.1 C-Test

Table 5 shows the means, difficulty values (P), standard deviations and reliabilities for the parallel forms of the C-Test.

Table 5

Means, standard deviations, difficulty levels, and reliability coefficients for the parallel C-Test forms

	test form	
	A	B
N	37	38
$\bar{x}$	60.2	67.4
S	20.3	19.7
P	50.2	56.2
r_{tt}	.78	.81

In view of the fact that both test versions were put together arbitrarily, that 4 new texts were used, and that no item analysis had been performed beforehand, the reliability coefficients are very satisfactory.

Both test forms have a medium level of difficulty for the group overall. The results are, however, not directly comparable because of the difference of 7 points between the means. For this reason a T-transformation was performed on both sets of scores, and these T-values were used for all analyses.

7.2 Equating of the school grades

It is customary in German to use an internal class-scale when determining school grades. Pupil performance is evaluated against the mean performance of the entire class, not against a criterion or some other absolute value. For this reason grades given in different classes are not comparable if the classes differ in their average performance level. Whether this is the case can be examined for instance by administering a test.

In this study precisely these conditions obtained: There were differences in performance level between the three classes, since the conception behind the German tripartite school system ist exactly that children with different types of ability (and thus, by implication, ability levels) should go to the different types of school.

For this reason I decided to choose the variable which correlated most highly with the school grades as an anchor, and to determine correction factors for the grades by regression procedures. The variable chosen was the total score on the DTD test. Table 6 shows the correction terms for the individual variables and school types.

Table 6

Correction terms for the grades

	German	grammar	English	Maths
Gymnasium	0	0	0	0
Realschule	0,4	0,5	0,6	0,6
Hauptschule	0,7	0,9	1,0	1,1

The corrected school grades were used for further analysis. The total score for DTD was ignored in subsequent calculations in order to avoid artefacts.

7.3 The correlation coefficients

In order to examine convergent and divergent validity for the C-Tests the correlations were calculated between the C-Test scores and all the other variables, both for the entire group and for the subgroups which took the different C-Test forms A and B. Table 7 shows the correlations with the individual variables ranked in order of magnitude.

Table 7

Correlations between C-Tests and the other variables

all		A		B	
DRT	.68	DRT	.63	TS	.76
TS	.66	PV	.62	DRT	.75
G German	.61	TS	.57	G German	.65
G gramm	.60	G German	.57	G gramm	.64
G Maths	.57	G gramm	.57	G Maths	.64
G Engl	.54	G Engl	.57	FDA	.57
IC	.54	IC	.54	G Engl	.54
FDA	.51	G Maths	.49	IC	.53
PV	.49	FDA	.46	RC	.50
FA	.44	FA	.42	FA	.46
RC	.43	AV	.39	AV	.40
AV	.40	RC	.36	PV	.36
D2	.20	D2	.25	D2	.14

$r_{SAB} = 0{,}63$ without PV $r'_{SAB} = 0{,}87$

We can draw the following conclusions from Table 7:

1. For the entire group the C-Test correlates most highly with the orthography test, with the grades for German and grammar, and with the subtest TS, which measures the ability to structure written linguistic material (= texts!). According to the authors of the DTD, TS reflects "a global performance which is necessary in different areas of language acquisition and mother tongue language proficiency". Here the correlations all reach at least the .6 level.
2. We find medium correlations (between .4 and .6) between C-Test scores and the school grade in English, the vocabulary and reading comprehension tests, and the tests which measure general intelligence and other intelligence factors (space, analogies, instruction comprehension).
3. There is no relationship to the concentration test.
4. The two subtests TS and IC in the upper part of the table constitute the part construct "successive extraction of information" according to the authors of the DTD test (see figure 1).

A comparison of the columns A and B in Table 7 shows very similar structures for each of them. An exception is the subtest Passive Vocabulary which correlates almost twice as high with the A form of the C-Test as with the B form.

If we calculate a rank correlation over the criterion variables between the two colums a value of .63 emerges if PV is included, and of .87 if this subtest is omitted.

A cautious interpretation of these results would be the following:

1. C-Tests measure chiefly general language proficiency, which is operationalised in the teacher ratings, the orthography test and the construct of successive extraction of information.
2. C-Tests also measure, though to a lesser degree, certain aspects of intelligence.
3. C-Test performance is independent of the ability to concentrate or of speed factors.
4. These results were confirmed in the cross-validation; The parallel test versions A and B are equivalent in their performance.

7.4 The factor analysis

In order to determine the factorial validity of the C-Tests intercorrelations between the variables both for the entire group and for each subgroup taking each of the two C-Test forms were calculated separately. The three factor analyses used the common factor method with iterative estimation of the communalities and a varimax rotation.

Factor extraction was broken off after the second factor, in spite of the fact that there were further factors with eigenvalues greater than 1 because:

- the proportion of variance explained diminished sharply
- the sample group was relatively small
- only two factors proved to be stable in the cross-validation study.

For the entire group the first two factors explained 61 % of the variance, for each of the subgroups 65 % and 61 % respectively.

Table 8 shows the loadings in factors 1 and 2 for the entire group after rotation. In order to bring out the factor structure more clearly those loadings have been marked which reach .6 or higher. The variables have been ordered according to the magnitudes of their loadings.

Table 8

Factor matrix for the entire group

variable	factor I	factor II
DRT	.87 o	-.04
TS	.80 o	.22
G gramm	.77 o	.32
G Engl	.74 o	.25
G German	.73 o	.27
C-Test	.71 o	.33
d2	-.11	.74 o
G Maths	.58	.65 o
FDA	.48	.63 o
IC	.43	.63 o
FA	.39	.61 o
RC	.49	.20
PV	.46	.26
AV	.23	.39

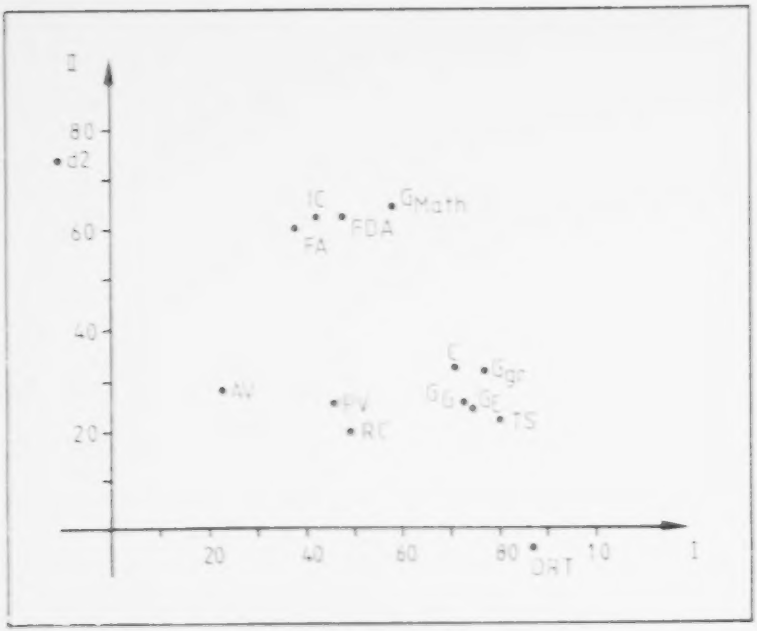
The first factor may be interpreted as a general language proficiency factor. This is supported by the fact that the dictation, which is a typical integrative language test from Spolsky's post-modern trend (Spolsky 1981), also loads high on this factor together with the school grades (also global measurements of language proficiency) and the subtest TS which was also described by the test authors as "global language performance". And the C-Test too has high loadings on the first factor.

The second factor has high loadings in the non-verbal intelligence test, in the maths grade, in the concentration test (which had a high speed component), and in the subtests instruction comprehension and analogies, which are, of course, subtests which are often found in other intelligence tests. This factor can be interpreted as a speeded factor of logical thinking. The C-Test has low loadings on this factor.

The factor matrix shows that it was not possible to achieve simple structure. This emerges very clearly in Figure 2, which is a graphic representation of the factor structure.

Figure 2

Factor structure for the entire group



Obviously the only way to achieve an ideal simple structure would be to select an oblique rotation. If the factor axes are permitted to be non-orthogonal, however, they have fairly high correlations with each other.

For the cross validation study a factor analysis of the same type was performed for each of the subgroups taking the parallel C-tests A and B. In Table 9 the factor matrices for the two groups are compared with each other. Here, only loadings of .6 and higher have been included.

Table 9

Factor matrices in the two subgroups

	A		B	
Variable	I	II	I	II
DRT	.92		.75	
TS	.87		.70	
G gramm	.80		.75	
G Engl	.73		.81	
G German	.66		.84	
C-Test	.66		.72	
d2		.71		(.45)
G Maths	(.43)	.72	.69	.62
FDA		.74		.60
IC		(.51)		.86
FA		.78		.60
RC				
PV				(-.02)
AV		.64		

If we take the small number of subjects into consideration, the similarity is remarkable. The only marked differences are

- in group B the maths grade has high loadings on both factors
- the subtest active vocabulary has a high loading on the intelligence factor for group A, but for group B it has no loading on either factor.

In view of the sample sizes these differences should not be interpreted.

8 Interpretation

This investigation is a pilot study, and therefore intended to be exploratory rather than conclusive. There are three reasons why the results must be interpreted with caution.

1. The sample size was small, particularly in the two subgroups.
2. As in any factor analytic study there is always the danger (cf. Raatz 1981) that the results are predetermined by the variables selected for inclusion. For instance it is quite within the bounds of possibility that no separate intelligence factor would have emerged if I had chosen to use a test of verbal intelligence, or if the concentration test had been omitted.

3. As a result of the equating of grades the true variance was increased, but very probably the estimation of the correction terms also increased the error variance. It is not possible to evaluate the effects of this on the results obtained.

However, provided we bear all these critical points in mind, we can still conclude that the examination of the factorial validity of C-Tests has produced similar results to all the validity studies performed on the C-Tests so far:
German C-Tests for L1-learners - native speakers of the language - measure primarily general language competence.

The close second order relationship between language proficiency and general intelligence, as shown in Figure 2, is very interesting, but really not surprising. There are empirical studies in psychology and linguistics with similar results, and sophisticated theories to explain them.

At this point, I shall only mention Oller (1978), who states: "language proficiency ... may account for the lion's share of variance in so-called IQ tests".

But the examination of this relationship was not the purpose of our study. Our aim was to investigate the validity of the C-Test, and here we have been relatively successful.

We plan to repeat the study with a larger sample taken from two different school grades. If this replication confirms the results presented here, then we will have very strong evidence for the validity of the C-Test. A further question in this planned investigation will be the relationship between C-Tests and IQ tests.

Brickenkamp, R. 1972. Test d2. Aufmerksamkeits-Belastungs-Test. Göttingen: Hogrefe.

Grotjahn, R. 1983. Entwicklung und Einsatz eines C-Tests "Französisch". Paper presented at the 14th Conference of the GAL, Duisburg, Sept. 1983.

Meis, R. 1969. DRT 4-5 Diagnostischer Rechtschreibtest für 4. und 5. Klassen. Weinheim, Berlin, Basel: Beltz.

Nauk, J. und Otte, Rolf. 1980. DTD Diagnostischer Test Deutsch. Braunschweig: Westermann.

Oller, J.W., Jr., 1978. "How important is language proficiency to IQ and other educational tests?". In Oller, J.W., Jr. and Perkins, K. (eds). 1978. Language in Education. Rowley, Mass.: Newbury House. 1-15.

Raatz, U. 1981. Sprachtheorie und Faktorenanalyse. Duisburg und Essen: Interuniversitäre Sprachtestgruppe IUS Workpapers 9.

Raatz, U. und Klein-Braley, C. 1983. "Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis". In Horn, R., Ingenkamp, K., Jäger, R.S. (Hrsg). 1983. Tests und Trends 3. Weinheim und Basel: Beltz. 107-138.

Spolsky, B. 1981. "Some ethical questions about language testing". In Klein-Braley, C. & Stevenson, D. (Eds). 1981. Practice and problems in language testing I. Frankfurt/M. und Bern: Lang. 5-21.

Wendeler, J. 1973. FDA 3-6 Frankfurter Denkaufgaben. Weinheim: Beltz.

The tests were supervised by Peter Klatte who performed the investigation as part of his examination dissertation. I should like to take this opportunity of thanking him.

LANGUAGE TESTS AS INDICATORS OF ACADEMIC ACHIEVEMENT*

Pauline M Kea (University of Dar es Salaam)

1 Introduction

As a result of concern over the falling English standards of students at the University of Dar es Salaam, where English is the medium of instruction, the Communication Skills Unit was established in 1978 with a view to improving the efficiency and effectiveness of student communication in English. Nonetheless, the extent to which English language proficiency is associated with academic performance is unclear and evidence from practical research into the predictive validity of language tests remains inconclusive (Spolsky 1967, Upshur 1967, Chai and Woehlke 1979, Wall 1981).

This paper is the result of work in progress on the concurrent and predictive validity of the University Screening Test (UST) and describes the way in which this measure of grammatical competence is related to academic performance.

2 Background

A number of studies investigating the relationship between language proficiency and academic achievement at university level have been conducted. There are several motivations for studies of this kind. Firstly, with large numbers of overseas students seeking admission at tertiary level, an indicator of language proficiency is frequently used as one of the criteria for placement. Secondly, considerable effort and resources have been spent developing language testing instruments used for purposes of university placement, eg TOEFL, EPTB, and more recently the ELTS (I). Thirdly, in the case of countries where there is a switch of medium of instruction from an indigenous language at primary and/or secondary level to English at secondary and/or tertiary level, there is increasing concern over declining standards of English language proficiency. Since it is assumed that language is an important factor in academic achievement, the language requirements for successful university studies and the provision of pre- and in-service language training for those who do not match up to these requirements are frequent topics of debate. It is this third reason which has been the primary motivation behind the present investigation.

* Financial assistance from the British Overseas Development Administration (ODA) to enable me to participate in the 7th International Language Testing Symposium of the IUS is gratefully acknowledged.

The first set of studies for review are those undertaken within a native English speaking context, and Chai and Woehlke (1979) provide a useful summary of results from a number of predictive validation studies in the United States. In their survey, the main language proficiency predictors were the TOEFL and the MTELP (2) but there was some variability in the criteria used for academic success: mid-year GPA, end of first year GPA, final year GPA. In all but one of the studies, predictive validity was estimated by a correlation between the test scores and the criterion performance. The levels of association between language proficiency and university achievement reported in the Chai and Woehlke survey show considerable variation both within and between the investigations cited. A total lack of dependence was observed between some of the language predictors and the criterion measures, in the work of, for example, Chase and Stallings (1966), Burguis and Naguil (1970), Abadzi (1975) and Hamlin (1976). On the other hand, Jones et al (1964) and Moran and Erion (1978) both report associations at the 5% level of significance, whilst Allen (1965), Hwang and Dizney (1970), and Gue and Holdaway (1973) report results with less than 1% probability that the observed association occurred by chance. In a more recent study, Homburg (1979) found no significant correlation between the TOEFL and GPA; neither did Wall (1981) who investigated the predictive validity of the Michigan Composition sub-test.

Research conducted in the United States suggests there is no systematic association between language proficiency and measures of academic achievement. In a few cases, there is a total lack of dependence between the two variables. In other studies positive correlations have been shown to exist but these are generally low. Chai and Woehlke (1979) further state that the maximum proportion of variance of the criteria predicted by the language test scores in their survey was in no case more than 50%, and in most cases it was less than 25%. Interpreting this as evidence of a weak association between predictor and criterion variables, they then questioned the commonly held assumption that academic success is dependent on language proficiency.

In the majority of studies carried out in Britain, estimates of predictive validity have not always been expressed in correlational terms but have made use of banding systems, or binary divisions such as 'pass' and 'fail'. Using both methods, Davies (1967) analysed the predictive and concurrent validity of English language reception tests on samples of overseas students, in the UK and overseas, and on a control group of native English speaking students, working at different academic levels. A range of validity coefficients were reported (+ .8 to - .2), some of which reached the 5% level of significance. Further, no support was found for the claim that language proficiency levels vary from one type of academic course to another: Davies identified a single cut off point which adequately accounted for academic achievement with nearly 80% success across academic levels. However, such a relationship was not demonstrated by the work of Sen (1970). In her study of 2300 overseas students studying in

Britain, in which a 'pass'/'fail' distinction on final GPA was matched with English language proficiency levels, she concluded that language ability was not predictive of academic success. On the other hand, Heaton and Pugh (1974) reported a significant difference between the mean English test scores of students who failed or passed their university courses. ($p < .005$ for undergraduates; $p < .01$ for postgraduates). Davies (1977) surveying 2000 students in Scotland obtained a 72% corroboration between EPTB scores and tutors' assessments of the 'adequacy' or 'inadequacy' of their students' language proficiency. Moller (1977) revealed a similar pattern of agreement (86%) between test scores and tutors' judgements, thus providing further evidence for the predictive validity of the EPTB in identifying students with language problems. James (1980) claimed predictive validity for the language tests at the University of Manchester in diagnosing students with linguistic deficiencies, reporting that two out of three students identified by their tests did in fact experience language difficulties. He reported that his study replicated exactly the findings of an earlier investigation by Jordan (3). Finally, Ingram (1973) in her work on the validation of ELBA (4) observed a significant relationship between language test scores and success in academic work, noting that the higher the score of ELBA the greater the chance of passing academic examinations.

In the British studies, there are no claims for a strong correlation between language proficiency and academic achievement. Although there is a significant relationship reported by Heaton and Pugh (1974), we may note that the highest amount of shared variance between the predictor and the criterion is just over 40%. On the other hand, these studies strongly suggest that students with poor language ability can be adequately screened by means of language tests and there is evidence that those thus identified stand a less than equal chance of succeeding in their studies.

The second set of studies to be reviewed here are those conducted in countries where English has the status of a second, third, or even fourth language and is used as the medium of instruction at tertiary level. At the University of Zambia, Heron (1970) examined the extent to which language training at university level contributed to the success of students as measured by mid-first year tests. He administered a series of English language tests, alongside reasoning tests, and controlled for different subject combinations. In the Humanities and Social Sciences, a systematic relationship between English performance and success in studies was observed in History and Political Science ($p < .01$), whereas no association with reasoning test results were identified. In the Natural Sciences, failure rates on a course in the 'Scientific Approach' were found to be dependent on reasoning ($p < .01$) and to a lesser extent on English proficiency ($p < .05$). A general tendency for Chemistry and Physics to be related to reasoning ability was observed (although the results were not significant), but no association with results on the English tests emerged. It was posited that

the overall lack of dependence between reasoning ability and mid-first year results, especially in the Humanities and Social Sciences, may have been attributable to the way in which staff were influenced by students' general ability to express their ideas in English. It was further suggested that the lack of dependence on English as far as the Science grades were concerned may have been a reflection of the type of examination used.

At the University of Haille Sellassie I in Ethiopia, King and King (1971) investigated nineteen variables in relation to academic performance and found that the best predictor of GPA, based on first semester results, was English language achievement as measured by the ESLCE (5) (r = .42). A very thorough correlational study by Douglas (1977) in the Sudan investigated the contributors to academic success at the University of Khartoum. The School Leaving Certificate results for English language were found to be important to university work, predicting academic achievement as well as anything else on the School Leaving Certificate Examination. It was further observed that English language proficiency was a major problem for first year students and that those demonstrating the highest level of English language competence were also those who were most successful in university examinations.

Five studies relevant to the concerns of this paper have been carried out in Tanzania. The first by Klingelhoffer (1966) investigated the relationship between High School Leaving Certificate results and composite first year grades in various subjects at the University of Dar es Salaam. The research failed to identify any systematic association between the predictor variables and the criterion measure of academic success. It was therefore suggested that the prediction of university achievement could be improved by including tests of English language proficiency, since success in studies was seen to be more likely a function of English language competence than any other single factor. However, in a study by Whitely (1969) correlations between first year education students' performance on a series of listening and comprehension tests and end of first year cumulative GPA scores were computed. No dependence was found between the language proficiency predictors and academic achievement. Since a report of this study is not available, it is not possible to explain the causes of this total lack of association.

Anderson's work (1975), also at the University of Dar es Salaam, was conducted as a complementary study to Whitely's (6), and examined students' productive language skills in written English in relation to educational performance, as measured by first year GPA. In contrast with all the studies previously mentioned, students' language competence was not measured through the administration of tests but through an analysis of language samples produced in naturalistic settings. Measures of sentence and clause length and complexity were computed, and the use of non-finite and passive verb forms were analysed. Some evidence for an association between these measures of English competence

and educational performance was obtained, although in general the correlations were not highly significant. It was observed that the measures of simplicity correlated most positively thus implying that students who wrote more simply (7) tended to be more academically successful in their first year at university.

In a more recent study in Tanzanian Secondary Schools, Mvungi (1982) examined the extent to which students who performed well in the medium of instruction (ie English) also performed well in their school subjects. Correlating English performance with History, Literature and Maths, significant results were reported between English language test scores and performance in the school subjects ($p < .01$). It is interesting to note that the level of significance is constant across the three school subjects. Computing the proportion of variance as predicted by English language performance we find that this falls in the region of 40%.

Data in a study by North (1983) provides evidence of a more consistent relationship between language abilities and academic achievement at the University of Dar es Salaam. Mid-year coursework marks for first year students in the Faculty of Medicine were compared with three measures of language ability and grades obtained on the Form VI School Leaving Certificate (on which university admissions are primarily based). The results indicated that all three language measures were strongly associated with coursework assessment scores, at the 1% level of significance and, in every case, that they correlated more highly with academic achievement than the Form VI qualifications. North's data therefore show language ability to be a fairly accurate and important predictor of academic success in first year studies. Additionally, the language tests in this investigation successfully diagnosed students with language problems with a satisfactory degree of accuracy. In a comparison of students scoring above and below half marks on the English language tests, only 3% of the high scorers had a failing average in coursework, as compared with 32% for the low scorers.

Taking an overview of the studies described in this paper, their results are open to various interpretations and, thus, the extent to which academic success is dependent on English language proficiency remains unclear. Several factors contribute to these inconclusive results. These include, for example, the considerable variation in the types of language tests used, and their proximity to the tests through which measures of academic achievement have been obtained. The characteristics of the samples are rarely held constant (cf Gue and Holdaway 1973), differing in size, nationality, subject specific orientation, level (undergraduate or postgraduate) and so on. The countries and the linguistic environments in which these investigations have been conducted also differ, and it has been further suggested (Kaplan and Jones 1964) that dual standards for the assessment of student performance are in operation, one for native English speaking students, another set for non-native speakers. The actual criterion measures of academic success

against which language proficiency is correlated, and the adequacy of these criterion measures, are also subject to variation. Further, there is diversity in both the methods used (correlational studies, banding or binary systems) in establishing the degree of association between predictor and criterion variables, and in the interpretation of results thus obtained.

It is not surprising, therefore, to find such varying degrees of association between language proficiency and academic performance because of the different circumstances and the variability in comprehensiveness of these studies. Additionally, as Douglas (1977) points out, contributors to academic success consist of a complex interaction of factors which include, importantly, general intellectual ability and learning ability. Language ability is therefore only one of a variety of factors.

Although the studies discussed above are not without theoretical interest, they have been primarily initiated for their potential usefulness in providing solutions to practical issues: their accuracy in selecting students for higher studies and/or screening students in need of additional language tuition. One of the main drawbacks of correlational studies is that they reflect the relative standing of students on different tests and are not directly influenced by differences in levels of attainment.

3 The Problem

At the University of Dar es Salaam, and elsewhere (8), there is very frequent mention of falling English standards. Interim measures to overcome the apparent problems have been implemented, and others are being sought (9), pending the strengthening of English language teaching in the secondary school sector.

The work reported in this paper has two related aims. Firstly, it examines the predictive validity of two language proficiency measures to determine academic success at first year university level. Secondly, it considers the accuracy of test scores on a test of grammatical competence to identify students likely to experience problems in their specialist courses.

4 Procedures

The predictors of academic success include two variables related to performance on the National School Leaving Certificate: Form IV English (now being considered as a compulsory requirement for university entrance) and Form VI results in students' subject areas which form the main basis for admission into the university. In addition to the O level English result, UST results are included, (as a measure of students' ability to apply their knowledge of English grammar in general areas of academic discourse).

Academic achievement is measured by the cumulative end of first year GPA, from which the results of all language measures have been excluded. The subjects considered here are first year students enrolled in four faculties (Commerce and Management (CM), Law (LL), Medicine (MD), and Science (SC)) in the 1982/83 academic session.

5 Results

A comparison of the three variables common to all university faculties is shown in Table 1 below.

Table 1

Comparison of Mean Scores and Standard Deviations

on Predictor Variables Common to all Faculties

VARIABLES		FACULTIES CM	LL	MD	SC
Entry Points	N	104	46	47	106
	$\bar{x}$	6.4	8.7	7.9	5.1
	σ	2.2	2.1	3.1	1.8
English O Level	N	120	57	55	102
	$\bar{x}$	2.6	3	3.3	2.5
	σ	.94	.91	.98	1.02
UST	N	104	54	53	122
	$\bar{x}$	63	68	71.6	64
	σ	16.3	17.6	16.5	13.7

The results indicate that the mean entry points for Law and Medicine tend to be the highest, and that there is a significant difference between the means of these two faculties with those of Science and Commerce and Management ($p < .001$). A similar pattern emerges with the O level English results, with higher mean scores for Law and Medicine. Again, a significant difference is observed between these faculties and those of Science and Commerce and Management ($p < .001$). In contrast, apart from a significant difference between Medicine and Science ($p < .01$), there is no evidence of significant variation in the mean UST scores for the samples analysed. This result is rather surprising because of the variation in failure rates obtained across faculties on the UST - 39% for Commerce and Management,

27% for Science, 17% for law, 11% for Medicine. This disparity in levels of grammatical competence is also confirmed by staff who teach students across university faculties (10). We may note, however, that the use of mean scores may obscure variance. Overall, these results suggest that the greatest difference in the students across the four faculties lies in their university entry points, although it should be noted that these points are not directly comparable as, for each faculty, they reflect a different combination of school subjects passed at Form VI level.

Table 2 below presents the correlation matrix for the predictor variables and GPA.

Table 2

Correlation Matrix for the Whole Sample (n = 364)

	English O Level	Entry Points	UST	GPA
English O Level	-	.19*	.07	.17*
Entry Points	.19*	-	.16*	.36**
UST	.07	.16*	-	.41**
GPA	.17*	.36**	.41**	-

* $p < .01$

** $p < .001$

These results show no relationship whatsoever between the UST and the O level English language result, whereas a low correlation is evident between the UST and entry points ($p < .01$). There is some indication that performance at O level in English is related to first year final examination results. The most statistically significant relationships exist between students' entry points and their UST scores with GPA ($p < .001$). However, this pattern is not consistently reflected in the breakdown of results per faculty presented next.

Tables 3 - 6 present the correlation matrices for the main predictor variables for the four faculties studied.

Table 3

Correlation Matrix for Medicine (N = 47)

	English O Level	Entry Points	UST	GPA
English O Level	-	.12	.38**(11)	.26
Entry Points	.12	-	-.05	.47**
UST	.38**	-.05	-	.54***
GPA	.26	.47**	.54***	-

Table 4

Correlation Matrix for Law (N = 41)

	English O Level	Entry Points	UST	GPA
English O Level	-	.24	.34*	.01
Entry Points	.24	-	.34*	.26
UST	.34*	.34*	-	.53***
GPA	.01	.26	.53***	-

Table 5

Correlation Matrix for Commerce and Management (N = 90)

	English O Level	Entry Points	UST	GPA
English O Level	-	.14	-.02	.29**
Entry Points	.14	-	.02	.23*
UST	-.02	.02	-	.47***
GPA	.29**	.23**	.47***	-

Table 6

Correlation Matrix for Science (N = 100)

	English O Level	Entry Points	UST	GPA
English O Level	-	-.07	.14	-.03
Entry Points	-.07	-	.07	.13
UST	.14	.07	-	.22*
GPA	-.03	.13	.22*	-

English language proficiency as demonstrated by the Form IV qualification appears unrelated to academic success, with the exception of the faculty of Commerce and Management ($p < .01$). No consistent relationship is observed between this last predictor and UST scores, although the association between the two are significant at $p < .05$ for students in Law and at $p < .01$ for students in Medicine. We may note from Table 1 that the mean O level results and UST scores are the highest in both Medicine and Law. On the other hand, it is perhaps not surprising for there to be a lack of systematic association between English O level and UST results not only because of the time lapse between the administrations of the two tests but also because a large proportion of university entrants do not proceed directly from school to university, many of whom will have been operating through the medium of Kiswahili (rather than English) in the intervening period.

With entry points and GPA, the results for the faculty samples do not indicate a consistent pattern. Varying degrees of significance ($p < .05$, $p < .001$) are indicated and, in the case of Science and Law, it appears that entry points are totally unrelated to first year success.

In every case, however, the UST proves to be the most consistent and highly significant predictor of academic success ($p < .001$ for Medicine, Law and Commerce and Management; $p < .05$ for Science). From the results presented in Tables 3 - 6 above, it appears that grammatical competence, as demonstrated by scores on the UST, is the best predictor of academic success at first year university level.

The next set of findings address the issue of the accuracy with which the UST identifies students who are likely to

underachieve in their university studies. The relationship between the UST scores and final first year results are presented in terms of a banding system. The underlying assumption of this analysis is that the lower the score on the UST, the greater the risk a student runs of failing in his area of subject specialisation. The results which follow in Table 7 are suggestive that this is the case.

Table 7

Comparison of UST Scores (MD, CM, SC) with First Year GPA Results (N = 268)

UST	GPA RESULTS	N
69% +	74% pass 26% fail	136
55 - 68%	46% pass 54% fail*	70
45 - 54%	38% pass 62% fail*	37
Below 44%	44% pass 56% fail*	25

The recommended cut off point for the UST, at present the same across faculties, is 54%. From Table 7, it can be seen that the level at which the proportions (12) start to reverse are at one banding level above the present cutting score. However, the trend is quite clear. The results provide evidence in support of the view that students with an inadequate grasp of English grammar are likely to be 'at risk' in their academic subjects, since the lower the score on the UST, the greater the likelihood of students supplementing (one or two subjects), repeating their first year, or being discontinued.

Additional evidence for the predictive validity of the UST is provided by the faculty breakdown given in Table 8. From these results, variation in the extent of failure rates at each level is to be noted. The appropriacy of a standard cutting score for all faculties is also an area for closer investigation. Nonetheless, the pattern emerging is substantially the same as for the whole sample.

Table 8

Comparison of UST Scores with First Year GPA Results (MD, CM, SC)

(N = 268)

UST	MD	N	CM	N	SC	N
69%+	71% pass 28.5% fail	49	89% pass 11% fail	45	62% pass 38% fail	42
55 - 68%	36% pass 64% fail*	22	71% pass 29% fail	21	33% pass 67% fail*	27
45 - 54%	33% pass 66% fail*	3	43% pass 57% fail*	21	31% pass 69% fail*	13
Below 44%	100% pass	1	57% pass 43% fail	14	20% pass 80% fail*	10
		75		101		92

Since the Communication Skills Unit also administers discipline-related tests designed to measure the academic study skills of notetaking from written and spoken inputs, reading comprehension, and writing, it would be informative to compare the results of these discourse processing tests with those of the UST. Of interest here would be the level of performance in academic study skills below which students stand a less than equal chance of succeeding at first year university level. With this information, the association between competences in grammar and academic communication skills could then be examined, thus providing useful input for those involved in the implementation and evaluation of remedial and developmental language training programmes.

6 Discussion

Two types of analysis have been presented here. In interpreting the results of the correlational analysis, it should be borne in mind that one would not necessarily expect a strong correlation between language performance and academic achievement to emerge. This is confirmed by the level of the association reported for the studies viewed at the beginning of this paper and by the present findings. Davies (1982) observes "it is well known that the correlation . . . with an academic criterion at the end of a year's course of study is typically about three". It could be, however, that the level of coefficient will vary according to the linguistic environment in which students undertake their academic course. The findings at the University of Hong Kong Language Centre (Low and Lee 1982, Fok 1982), and in Dar es Salaam (North 1983), together with those presented in this paper are suggestive, but by no means conclusive, that there is a greater dependence between language proficiency and academic success in ESL rather than in native English speaking contexts. Further, it may be noted that the language measures in the studies just cited have been specifically developed as tests of English proficiency for academic purposes, following a less traditional and more communicative approach to testing. In all of these tests, the communicative demands more closely approximate those required by students' academic courses, thus it may be the case that an English test will predict the results of another test in English if it is similar in form (Anderson 1975). This aspect of test validity has been raised elsewhere (Chai and Woehkle 1979, Wall 1981).

The issue, I suggest, is not really one of dependence, in correlational terms, of academic success on language proficiency, but rather one of the actual predictive potential of English tests to accurately screen out those students 'at risk' (without further language training) when following courses in their area of subject specialisation. The use of expectancy tables provides a more immediate and useful illustration of the extent of the relationship between success in studies and language proficiency, and of the power of the latter to predict criterion performance. According to the data presented, students who score below a certain level on the UST stand a greater chance of failing in their academic subjects than those who score above this level. This supports the hypothesis underlying the introduction of the UST at the University of Dar es Salaam, and is evidence of its predictive power.

It was observed earlier that entry points required for university admission vary from faculty to faculty, but a comparison of mean scores on the UST across faculties revealed little variation. Further analyses of all university faculties is required for these variables, and the possible existence of varying language proficiency threshold levels for students in different faculties, or clusters of faculties, should be examined.

Subsequent investigations will necessarily raise several interesting issues. These include the relationship between alternative test types (skill-based, grammar-based) and test formats (multiple choice, cloze, guided production tasks) and their independent or cumulative contributions to academic success. The influence of test content on the performance of students in different subject areas is a very relevant concern (13), especially with the increase in number of ESP testing modules (14).

7 Conclusion

This investigation has been exploratory, arising primarily from the need to validate the UST at the University of Dar es Salaam. Subsequently, in response to a more recent request for guidance on the status of proven proficiency in English as a prerequisite for university admissions, it has addressed the issue of contributors to academic success. In respect of the former objective, this paper has considered evidence for the predictive validity of the UST and it appears that this measure is efficient in distinguishing between two categories of students. Concurrent validity data, reported elsewhere (Rea 1983), strongly suggests that students who fall below the specified threshold level on the UST are in need of remedial grammar work. The results reported here indicate that these same students are also very likely to encounter difficulties in their specialist subjects.

The second objective is considerably more problematic. I have attempted to show the influence of proficiency in the medium of instruction on academic performance. There is an indication that the introduction of a prerequisite pass at English O level will not significantly alter the status quo. A viable alternative to this might be the administration of the UST, or one of the other language tests being analysed, in their present, or modified, forms.

An important factor emerging from studies in this area is the variation in results within the same institution, from institution to institution, and from country to country. It is clear that any policy decisions on compulsory language requirements for university entrance will depend on the findings of a comprehensive study. By the very nature of the complexity of the multilingual environment in Tanzania, and the complex of factors inherent within the predictor and criterion variables of academic success, future analyses should systematically address potential sources of variance. Further, the effects of language training provided, in the intervening period between initial language testing and the point at which data on the criterion variable of academic achievement are obtained, should be taken into consideration. Because of the dynamic nature of the linguistic situation in Tanzania, data over several years should be examined.

Finally, test scores are not perfect measures and they should not, therefore, be treated as such. They exist to provide guidance and, following the advice in the TOEFL Manual 1981, each institution is well advised to match the suitability of English test scores with its own specific purposes for testing. Because of the waste of national resources in the case of students who fail to complete their university education (North 1983) and the time and resources spent in developing and maintaining language training programmes, detailed investigations in countries in similar circumstances to Tanzania would not seem to be misguided. At one and the same time, they will provide insights into theoretical issues of current concern to language testers and they will also signal the way ahead for useful developments in the formulation of language policies and the development of language training programmes from which the majority may derive maximum benefit.

Notes

1 The Test of English as a Foreign Language; the English Proficiency Battery, also known as the Davies Test; the ELTS has been devised as a replacement for the EPTB and is administered by the British Council in association with the Cambridge Examinations Syndicate.

2 The Michigan Test of English Language Proficiency.

3 Jordan, B 1976. Cited in James, K 1980.

4 The English Language Test Battery developed in the Department of Linguistics at the University of Edinburgh.

5 The Ethiopian School Leaving Certificate Examination.

6 The only references to Whitely's research are available in Anderson 1975.

7 Students who used shorted T units and fewer subordinate and embedded clauses.

8 At the University of Makerere, Uganda; the English Language Unit at the University of Zimbabwe; the call for concerted efforts to improve the teaching of English at secondary school level in Botswana, to mention a few.

9 For instance, the introduction of a pass at English O level as a prerequisite for university admission.

10 That is, CSU staff who teach the Intensive Grammar Programme in all the university faculties.

11 The significance levels for all tables are as follows:
* $p < .05$; ** $p < .01$; *** $p < .001$.

12 Pass comprises A, B, and C grades on GPA; Fail covers D and E grades and also students who supplement one or two subjects, repeat the year, or are discontinued from university studies.

13 See J C Anderson and A H Urquhart in this volume; also their paper "The Effect of Student Background Discipline on Comprehension: a pilot study", in Developments in Language Testing, edited by D Porter and A Hughes, London: Academic Press 1983.

14 Particularly in connection with the ELTS which is administered worldwide.

Bibliography

Abadzi, H 1975. "Evaluation of foreign students admissions procedures used at the University of Alabama". Cited in Chai, S H and Woehkle, P L 1979: 20.

Allen, W P 1965. "International student achievement: English test scores related to first semester grades". Cited in Chai, S H and Woehlke, P L 1979: 20.

Anderson, R 1975. "The language factor in educational performance at university level". Journal of the Language Association of Eastern Africa, Vol 3, No 2: 138-163.

Burguis, T C and Naguil, A F 1970. English Language Proficiency and Academic Achievement among Students of English as a Second Language at the College Level. Cited in Chai, S H and Woehkle, P L 1979:20.

Chai, S H and Woehlke, P L 1979. "The predictive ability of standardized tests of English as a foreign language". Occasional Papers on Linguistics 6: 15-28. Department of Linguistics, Southern Illinois University at Carbondale: 15-28.

Chase, C I and Stallings, W M 1966. "Tests of English language as predictors of success for foreign students". Monographs of the Bureau of Educational Studies, Indiana Studies in Prediction 8: 1-24.

Davies, A 1967. "The English proficiency of overseas students". British Journal of Educational Psychology 37, part 2: 165-174.

Davies, A 1977. "Do foreign students have problems?" In Cowie, A P and Heaton, J B (eds) 1977. English for Academic Purposes. BAAL/Selmous Publication: 34-36.

Davies, A 1982. "ELTS Validation Project". In Hughes, A and Porter, D (eds) 1982. Language Testing Newsletter 2, Spring: 5.

Douglas, D 1977. From School to University, The Study Habits Research Project Final Report. Khartoum: Khartoum University Press.

Fok, A 1982. "Language proficiency and its related factors". Paper presented at the International Symposium on Language Testing, University of Hong Kong, December 1982.

Gue, L R and Holdaway, E A 1973. "English proficiency tests as predictors of success in graduate studies in Education". Language Learning 3: 89-103.

Hamlin, E C 1976. "An analysis of the relationships between the English language proficiency scores of entering foreign graduate students and their academic achievement in an advanced degree program". Cited in Chai, S H and Woehlke, P L 1979: 21.

Heaton, J B and Pugh, A K 1974. A Study of the Relationship between Scores Obtained by Overseas Students on a Test of English Proficiency and the Examination Results in Their University Courses. University of Leeds.

Heron, A 1970. "The relative importance of reasoning ability and of English vocabulary and comprehension in first year university achievement by Zambian students". In Gorman, T P (ed) 1970. Language in Education in Eastern Africa. Nairobi: Oxford University Press: 114-128.

Homburg, T J 1979. "TOEFL and GPA: An Analysis of Correlations". Occasional Papers on Linguistics 6: 29-35. Department of Linguistics, Southern Illinois University at Carboudale.

Hwang, K Y and Dizney, H F 1970. "Predictive validity of the Test of English as a Foreign Language for Chinese Graduate Students at an American University". Educational and Psychological Measurement 30, No 2: 475-477.

Ingram, E 1973. "English standards of foreign students". University of Edinburgh Bulletin 9, No 12: 4-5.

James, K 1980. "Survey of University of Manchester overseas post-graduate students' initial level of competence in English and their subsequent academic performance: The British Council: 126-133.

Jones, R A, Kaplan, R and Michael, W B 1964. "The predictive validity of a modified battery of tests in language skills for foreign students at an American University". Educational and Psychological Measurement 24: 961-965.

Kaplan, R B and Jones, R A 1964. "Evaluation of relative foreign student success". Language Learning 14, Nos 3 - 4.

Klingelhoffer, E L 1967. Studies of Tanzanian Students. Dar es Salaam: University College.

King, M and King J. "Some correlates of university performance in a developing country: the case of Ethiopia". Journal of Cross-Cultural Psychology 2, No 3: 293-300.

Low, G D and Lee, Y P. "How shall a test be referenced?" International Symposium on Language Testing, University of Hong Kong, December 1982.

Moller, A 1977. "A case for a crude test overseas". In Cowie, A P and Heaton, J B (eds) 1977. English for Academic Purposes. BAAL/Selmous Publication: 25-33.

Moran, R T and Erion, J G 1973. "Predictive validity of the CELT". TESL Reporter 23: 89-103.

Mvungi, M 1982. Relationship Between Performance in Instructional Medium and Performance in Selected Subjects in Tanzanian Secondary Schools. PhD Thesis, University of Dar es Salaam.

North, S 1983. "Language proficiency and first year university achievement: a case study at the University of Dar es Salaam". Communication Skills Unit, University of Dar es Salaam.

Rea, P M 1983. "Linguistic competence and academic performance: the University Screening Test". Language for Education 2: 4-15 Communication Skills Unit: University of Dar es Salaam.

Sen, A 1970. Problems of Overseas Students and Nurses. London: National Foundation for Educational Research.

Spolsky, B 1967. "Do they know Enough English? Some notes on the problems of assessing the proficiency in English of foreign students". Selected in NAFSA 1967, Conference Papers of the Association of English as a Second Language. New York: National Association for Foreign Student Affairs.

Upshur, J A 1967. "English language tests and prediction of academic success". In NAFSA ibid.

Wall, D 1981. "A study of the predictive validity of the Michigan Battery, with special reference to the composition component". In Culhane, T, Klein Braley, C and Stevenson, D K (eds) 1981. Practice and Problems in Language Testing. University of Essex, Occasional Papers 26: 156-169.

INPUT AND OUTPUT IN LANGUAGE TESTING

Elana Shohamy (Tel Aviv University)

1. Introduction

The issue of quality tests has been discussed talked about, and written about for some time. Stevenson (1981) provides a distinction between "good" and "bad" language tests by pointing to two views of language testing - a "restricted" view which focuses on the technical aspects of tests, and an "expanded" view which focuses on the scientific aspect of such tests. He elaborates on those two views by stating that the restricted view refers to situations where the test constructors have little or not training and expertise in testing and measurement theories. They are only marginally concerned with aspects of validity, reliability and language testing theories. Furthermore, testing is perceived as a technical skill and is concerned mainly with the "how to do it". It focuses only on the visible content, construction and scoring of tests.

The "expanded" view, on the other hand, refers to a situation where there is awareness of a scientific approach to the study of language behaviour and measurement of language. Those who have an expanded view to testing are interested in various testing approaches such as essays, sociolinguistic elicitation procedures and their effects. Moreover, they are aware of the controversies regarding the nature of language and what is actually being measured when language ability is tested. The expanded view is therefore concerned with basic questions of validity and validation. It is less assured, less prescriptive and much more aware of the great complexity of what it is attempting to measure. It is based on a theory of what it means to know a language. Because of the incompleteness of such a theory it advocates the use of multiple procedures for language testing. Its prime concern is construct validity, while reliability is taken for granted. It asks itself "why test" and "what" exactly can and should be tested. The expanded view then calls for the development of testing instruments which are based on a clear definition of the domain of second language learning and what it means to know a language.

Stevenson (1982) claims that the restricted view is narrow since the emphasis is on the "how" of testing, that is, the techniques found in text books for constructing tests which are much less important than an understanding of the why and the what. Unfortunately, he claims, it is frequently the how that receives most of the attention from language teachers. Yet, any language test is necessarily dependent upon the purpose of the testing procedure, and upon how one answers the question of what it means to know a language.

This paper will provide a different distinction between two approaches to language testing where each approach is based upon a

different assumption - the synthesis of both approaches can result in quality language tests. It is claimed that in the field of language testing today, each of these approaches is used independently while the production of good tests requires the integration of both.

2. The Problem

In the past few years there have been serious discussions in the field of language testing about measurement issues such as test reliability, and validity, and little mention, if at all, on issues related to language learning and acquisition as they relate to language testing. Measurement experts are used to dealing with psychometric issues without finding a need to delve into the theory of the domain being tested. After all psychometrics is a well established field which devotes itself to such issues and there is an understanding that measurement experts do not need to know much about the domain being tested.

However, foreign language learning is different than other subjects. It is unique in several aspects: first, unlike other subjects it can be learned witnout formal teaching, thus in constructing tests there is not always a need to relate to and depend on a specific curriculum for deciding what to test. Second, unlike other school subjects, the process of second language learning as well as language proficiency have been researched and there are some findings in areas such as the successful language learner, attainable levels of proficiency, effects of social and individual factors on learning a foreign language, all of which have an impact on the testing of foreign languages. Thus, the language tester needs to know about the field of Applied Linguistics, the definitions of language and its knowledge, the importance of knowing how to use and communicate in a language, and that knowing a language is perceived as something more than discrete elements of lexicon and grammar.

Thus, even a well-trained measurement expert who is knowledgeable in modern and classical measurement theories in areas such as item analysis, reliability, validity and latent trait models will not be able to construct good language tests without knowing what it means to know a language. He needs to know what the various phases of language learning are and what the expected levels of foreign language proficiency at different stages are. Now knowing about these aspects makes it impossible for the measurement expert to construct tests which will be construct valid since he cannot know what should be tested, and more important, what the criteria for accuracy and correctness are. He may be testing the "wrong" things and may be scoring the tests based on the "wrong" criteria. One example to illustrate this phenomenon is the testing of pronunciation, or accent in foreign languages. For many years accent was tested in various language tests in a variety of ways. However, research of the past few years in second language acquisition has shown that only a very small portion of foreign language learners can fully acquire the foreign language accent and that there are various degrees of

accent proficiency that are attainable by the foreign language learner, depending on specific situations. Such information is essential for the tester to have in order to construct valid speaking tests. Still, this type of information by itself is clearly necessary but not sufficient for producing good tests of speaking. There is also a need to impose rigorous psychometric criteria on that information to convert it to a test. The test needs to differentiate and discriminate among a wide variety of speakers, all the items must relate to the construct being tested, it must give information on the likelihood of obtaining similar results upon a second administration of the same test, and it must be difficult enough for the good students and easy enough for low level students. It is clear from this example that the construction of quality language tests requires the integration of knowledge from these two domains - language learning and measurement theories.

Another example to illustrate the same point is the testing of sociolinguistic proficiency. Till today, sociolinguistic proficiency has been rarely tested. Most language tests still focus on linguistic aspects as the major criteria for test construction. However, the past few years have brought about an awareness of the use of language within a social context. It is apparent today that knowing a language means also knowing when, where, how and to whom to use it. Overlooking the state of the art in what knowing a language means today, implies the construction of tests which fail to tap the construct of language proficiency in its full and complete definition. For the person who is aware of sociolinguistic definitions, the construction of such tests will obviously involve imposing rigorous measurement criteria to convert this sociolinguistic information into tests of sociolinguistic proficiency.

Unfortunately, many of the tests used for arriving at conclusions about individuals are based on instruments which are incomplete either because they are not based on a fuller theory of knowing a language or they do not possess the necessary measurement properties. Thus, tests which are based on language learning theories without good psychometric theories are not sufficient in the same way that tests which possess measurement properties but are not based on complete language learning theories are insufficient.

The claim that is made here is that there is a need for greater reliance on and application of existing theories of language learning in order to define more fully the domain being tested. Any other way is restricted and narrow in its scope and inevitably leads to tests which lack construct validity. At the same time there is a need to impose rigorous measurement criteria on tasks which draw on language learning theories. Without these criteria this knowledge cannot serve as a valid basis for testing instruments. Thus knowledge of theories of second language learning on one hand and measurement theories on the other, are necessary but in themselves insufficient to produce quality language tests - there is a need for integration of the two domains.

3. Input and Output

The two domains - language learning and measurement have different, yet complementary roles and purposes in producing language tests. Language learning theories can bring the INPUT component into language testing i.e. defining what should be tested and what the appropriate language behaviours are. Measurement theories on the other hand can bring the OUTPUT component of the test i.e. how can language performance and proficiency turn into tests of these constructs.

3.1 The input component

Input in second language is concerned with the definition of language and what it means to know a language for given purposes. Second language test constructors must define proficiency levels as objectives or as standards by which to assess foreign language knowledge. The objectives or standards can serve as criteria for judgement.

Cummins and Stern (1981) provide a comprehensive view of some approaches to determine what it means to know a language. They distinguish among four distinct approaches to the phenomenon and claim that none of the four is really sufficient since there is a need of all the four to be combined for a fuller picture of language proficiency. The four approaches are: a) theoretical concepts; b) rating scale; c) formal proficiency tests; and d) inter-language research.

The four approaches will be briefly reviewed here with the understanding that each, and perhaps all four combined, could serve as a basis for defining the domain that can be used as the input component of a test.

The first approach is a theoretically based concept of proficiency. Three concepts are described here: a) proficiency defined in linguistic content. Up till the 1970's the theoretical linguistic model included phonology, vocabulary and grammar. More recently features of semantics, discourse and sociolinguistics have been added to the theoretical concept of proficiency. Today the concept of proficiency is expressed in communicative and not only in grammatical terms, but obviously do not exclude grammar; b) the second concept is the psychological or behavioural one. Such description of proficiency ranges from the abstract concept of "competence" to the more concrete term of listening, speaking, reading, and writing skills. At the most concrete level, proficiency is described as categories of language items, situations, speech functions and topics. Others define proficiency in terms of syllabus of items for specific groups of learners. c) A third concept that is made up of both linguistic and behavioural features is that of stages or levels of L-2 proficiency, which describes schemes involving different levels, each described in terms of the four language skills.

The second approach to proficiency is through describing

different proficiency levels by means of a rating scale. This approach is made on the basis of learner's practical knowledge at different stages. The FSI and Defense Language Institute have designed rating scales for the four language skills, which are divided in terms of communicative functional skills. Such rating scales describe expected standards for achievement levels in terms of the objectives to be aimed at. Stern and Cummins claim that rating scales can provide useful impressionistic descriptions of the stages and development of proficiency at different levels.

The third approach defines proficiency as a result of formal tests. Language proficiency tests imply a conception of proficiency that is based on empirical data. However, since researchers still have no definite answer to the question of the nature of language proficiency, some language tests results can do that. Unfortunately, they claim, most present language tests primarily measure the academically related aspect of proficiency and not the communicative and interpersonal components.

The fourth approach to proficiency is provided by the inter-language studies where the researcher examines closely what learners actually do in their L-2 performance. They refer to Selinker's work which describes the proficiency of L-2 learners as a linguistic system in its own right between the L-1 and the target language. The contribution of interlanguage research to this description of proficiency may lay in the concrete data that such studies provide on the development of proficiency. Such data can be extremely useful for providing a basis upon which one can plan and build tests. The work of Blum-Kulka and Olshtain (1984) in which they describe inter-language of speech acts of apologies and requests can also be a promising description of sociolinguistic data that can be extremely useful in constructing tests based on sociolinguistic input.

It is clear from the review above that there are various sources that language testers can draw upon in order to obtain the necessary input data. By focusing on the input aspects, the language tester can be confident that he is building on a solid foundation.

3.2 The output component

Input data is the foundation of a test but in itself is not sufficient for producing good tests. There is a need to convert the information described in the input into a test by providing empirical measurement data. The input data can serve as a basis for the language test which also needs to possess the appropriate measurement properties. These properties relate to the accuracy of the tasks and their ability to provide good judgement. The kind of measurement criteria imposed depend on the purpose of the tests and thus will differ from one test to another. However, all the criteria are intended to indicate that the testing instrument is capable of judging and making decisions about individuals' language knowledge. The main properties generally used for the output are the various types of reliability and validity as well as item analysis. It is suggested that the measurement criteria be imposed after the input has been defined.

3.3 The integration

The language tester's role is to integrate the input and the output components. The language tester needs to construct testing tasks based on the input, that lend themselves to testing and that will also possess the needed measurement properties.

4. Real-life examples

The type of work done in language testing in the past few years and some of the papers presented at this meeting, can serve as live examples to illustrate the points made above about input and output in language testing. Two types of tests seem to emerge in the field - input type tests and output tests. The input type tests are concerned primarily with the need to base tests on theories of language proficiency and with defining language as communicative interactions. The example to illustrate this is the entrance exam to British Universities (the TEAP test) as it was described in this meeting by Weir and Emmett. This test is based on a careful analysis of language performances that students are likely to engage in when studying at British universities. Careful descriptions of these tasks are provided as the input aspect of the test. Clearly, this approach is very different from the traditional approach where tests are constructed without this input foundation, usually by arbitrary decisions, and often just as results of correlations with other existing and usually bad tests. The input approach provides a careful analysis of language proficiency for the given purpose of testing. The tasks described in the TEAP are drawn from actual language performance, however, it is clear that these type of tasks by themselves cannot provide the necessary testing data until it has been proven that they also possess the necessary measurement criteria and requirements. Very little information is available to date on the output aspect of that test.

The paper presented by Baghdady in this meeting provides another example to prove that point. In his test there is also a very strong input component based on a systematic analysis and description of language proficiency, but there too what is lacking is the transformation of these tasks, most of which can qualify as good teaching tasks, into testing tasks.

The other type of tests that emerge are the output tests, which focus primarily on measurement properties. These tests are able to discriminate among various proficiency levels what they may be, they have high concurrent validity with other tests and high reliability. Tests such as the C-test which was described in this meeting, and in other IUS meetings by Klein-Braley and Raatz can be classified as output type of tests. The C-test is a very powerful instrument from a measurement point of view, it has excellent statistical properties - it discriminates among heterogeneous and homogeneous groups, between native and non-native speakers, among learners of different ages and levels and it has high inter-consistency reliability. And yet there

is no language learning theory to explain what it is doing and more importantly, why. The lack of language proficiency basis, the input, to serve as the foundation for the powerful output of the test, is definitely a limitation of that test. Supplying the seocnd half of every second word must be justified theoretically in terms of language learning. This is difficult to accept for these who share an expanded view of language testing.

In summary, these examples attempt to illustrate the two types of tests - on one hand tests which are strong on the input side and are based on an analysis of language proficiency but fail to prove their strength from a measurement point of view, and on the other hand, tests which are strong on the output side - pass various measurement criteria, but fail to provide a theory to explain these results, and therefore it is not really clear what they are measuring.

5. Conclusions

This paper criticized the approaches currently used for the development of language tests where some tests are based on a strong input aspect - what goes into a language test, and others are based on the output component - providing appropriate measurement criteria. It advocates the integration of both components, input and output, in order to produce good language tests. The paper recommends the descriptions provided by Stern and Cummins as a point of reference to focus on the input, while the output can be drawn from the current measurement theories. Needless to say, many of the tests which are available today do not possess either of these components.

It should be noted, however, that theories of language proficiency are still not complete and many of the sources of the input component have only basic preliminary information that could be used for test construction. At the same time measurement theories also still fail to provide sufficient and full direction regarding the psychometric procedures which should be used for dealing with language as it is defined today - communicative and authentic. In future years special efforts should be devoted to advancing these two components, input and output, in an integrated way into an "expanded" theory of language testing.

Bibliography

Blum-Kulka S. and E. Olshtain, 1984. Requests and Apologies: A Cross-Cultural Study of Speech-Act Realization Pattern, to appear in Applied Linguistics, Vol.5, No. 3

Stern H. and J. Cummins, Language Teaching/Language Research, 1980. In ACTFL Annual Review - Action for the 80s. National Textbook Company, Chicago, Illinois.

Stevenson Douglas K. Language, 1981. "Testing and Academic Accountability: On Redefining the Role of Language Testing in Language Teaching", IRAL, Vol. IX/1

Stevenson Douglas K., 1982. "Language Testing - All of the Above" in James C. (ed.) Research (working title) ACTFL Annual Review, Vol. 14, National Textbook Company, Skokie, Ill.

LANGUAGE TESTING WITH IMMIGRANT CHILDREN

Edgar Sußmilch (University of Düsseldorf)

1 The educational situation of immigrant children in West Germany

The number of non-German pupils in the Federal Republic of Germany and West-Berlin rose steeply in the Seventies. Recruitment of foreign workers ceased in 1973 but many children have since entered the country to join their parents. In addition many children of immigrant workers have been born in Germany. In the school year 1981-82 816 000 non-German children were attending schools in the Federal Republic. In the Grundschulen (primary schools) and the Hauptschulen (secondary modern schools) the proportion of non-Germans reached 12.6% overall, in North-Rhine-Westphalia 15.3%, and in Berlin, which has the highest proportion of immigrant pupils, every fourth child in the non-selective schools was a foreigner. However, because of the unequal distribution over the country in some cities and city areas the proportions are even higher.

Only recently and very gradually have the individual federal states begun to develop concepts and strategies to improve the educational provisions for these children and adolescents, and one of the undecided issues is the relative importance of the native language and of German as a second language. It is obvious that German is a necessary precondition for success in school and employment opportunities, as well as for social and cultural integration in general. But teaching German under these conditions is a very difficult enterprise. First of all, the teachers have not been trained to perform this task. There are few suitable text books available. Classes are large, and the non-German children vary greatly in their command of the German language. However, since remedial language teaching must take this heterogenity as its starting point, an objective technique for determining proficiency levels would be useful.

2 Measuring proficiency in German as a second language

2.1 Language diagnosis and language training

Tests in the school setting are instruments for acquiring information to be used in planning teaching strategies and evaluating teaching success. They have an important contribution to make as objective assessment instruments to help the teacher in making judgements about his pupils. For the immigrant children tests could support the teacher in making classification decisions on school entry as well as in determining whether pupils should be assigned to preparatory or to normal classes. A further decision point where language tests could be useful is at the end of the fourth grade where pupils can be sent either to a selective secondary school or to the Hauptschule. Language tests can also be used for diagnosis and treatment decisions.

In the last few years a number of test development projects have been undertaken. They have a variety of different linguistic foundations and pedagogical objectives. Because there are at present virtually no suitable tests commercially available, many approaches have been tried by practising teachers in order to provide a better basis for evaluating the German proficiency of the immigrant children and for developing appropriate teaching strategies. Often the aim has been to combine diagnosis and therapy. However, in many cases these tests do not proceed beyond initial trials, and empirical investigation and analysis are not performed so that such procedures remain questionable both linguistically and psychometrically. A brief survey of the available approaches will show this more clearly.

2.2 Informal language tests

Ihssen, speaking about the methodological problems of language diagnosis with immigrant children, demands "an account of the individual language biography and the sociolinguistic conditions ... in order to obtain an adequate assessment of the developmental level in the first and second language, for parental counselling, for the planning, implementation and evaluation of supporting measures" (1980:40). As the primary source of information for a diagnosis of language development he proposes the analysis of spontaneous language behaviour in a natural communication situation. To do this he takes a sample (representative if possible) of 50 consecutive sentences. Using a grid taken from child language development research he analyses word and sentence structure (morphology and syntax). His categories include, for instance, the average number of morphemes per sentence, wrongly used grammatical morphemes and word order errors.

Many other approaches follow Ihssen's lead. Informal tests are developed which cover only a limited segment of language derived from a structural analysis. Obviously it is felt that such approaches are particularly suited for therapeutic purposes. Standardised (group) tests which comply with the normal test requirements of objectivity, reliability and validity are rejected, since it is felt that they are incompatible with the therapeutic approach. Instead tests are developed which cannot be evaluated by the traditional techniques and which, therefore, are of dubious value.

Scherzinger and Scherzinger (1981) expand Ihssen's approach by using picture stories as elicitation techniques. This permits the collection of comparable data for each subject. The same picture sequence is used to elicit an oral and a written sample. Here again only morphosyntactical features are scored, although qualitative criteria, such as breadth of vocabulary and communicative competence, are mentioned.

Other test projects also stress the analysis of morphosyntactical features in spontaneous speech samples: obviously it is felt that a structural model of language is particularly suitable for the assessment of L2 proficiency with immigrant children (cf. Portz and Pfaff 1981; Steinmuller 1983). A technique for analysing written texts is offered by Hegele (1981). The aim of her investigation is to provide baseline data for planning therapeutic measures and for

evaluating their effectivity.

In three further projects the attempt has been made to provide extensive item pools and to provide information about test administration and scoring. However, even these are informal tests since they do not conform to the accepted test standards, which reduces their usefulness. In addition item analyses have not been attempted so that improvement of the individual items is not possible.

In the first project the test for transfer into the normal 4th grade class was produced by the Akademie fur Lehrerfortbildung Dillingen (no date). It was the first extensive language test developed for immigrant children. It consists of an individual interview with tests of free speech production, guided speech production, listening comprehension, reading comprehension and a written group test with the subtests vocabulary, grammatical structure and dictation. No details are provided about the total time needed for test administration. A first trial with 34 subjects showed a wide range of scores.

Luchtenberg, Neumann, and Wespel (1982) also developed an informal test. It is restricted to oral communicative ability: understanding simple instructions, simple questions and statements, active vocabulary, simple word and sentence structures in German. For this individual test (which can take up to two hours to administer) the only data reported consist of a fictitious results table for 5 subjects.

Fliegner and Gogolin (1982) provide a test to be used with school beginners. Revised material for the areas listening comprehension and speaking, split into word, sentence and text level, is accompanied by a teachers' manual. The materials provided form the basis for a 20 minute individual interview. However, no controlled trials have been conducted and the revision was performed on the basis of critical suggestions and teacher reports. An empirical investigation of the materials is planned and will be conducted by Boos-Nunning in Essen.

2.3 Standardised language tests

In a number of studies language tests developed for German L1 speakers have been administered to immigrant children. Mitzschke (1978) used the reading comprehension test VL 5-6 for a group of non-German youngsters involved in pre-apprenticeship training (N = 37). Damanakis (1978) also used available German tests: the subtests Vocabulary and Reading Comprehension from the AST 2 for Greek pupils (N = 49) in preparatory classes in the second grade. In a second study, concerned with L1 and L2 development in Greek pupils attending normal classes in the 4th and 6th grade (N = 114), Damanakis used the vocabulary test WST 4-6 and the orthography test RST 4+ together with two Greek tests he developed himself. Both he and Mitzschke compared their results with those of the German norming samples, although the comparison is restricted to means and standard deviations.

Biehl (1982) has reported the results of a pilot study performed

in the framework of a Duisburg project investigating attained language proficiency and conditions of language learning. Tape recordings were made of an interview and of a picture story narrative, and newly developed vocabulary and syntax tests were administered. For the vocabulary test 60 words were selected from a 1 000-word basic vocabulary list and L1 translations were combined with L1 distractors to produce multiple choice items. The syntax test consists of pictures accompanied by sentences of varying syntactical complexity. The subject has to state who is doing what in the pictures, and in order to assign the correct names he must understand the accompanying sentence. Yet again, for this study involving 75 subjects no data about test analysis are available, which is rather surprising in view of the fact that Duisburg is known to have its fair share of experts in the field of language testing. Interesting relationships between the language data and the social data have been reported by Biehl, but since not enough information has been presented about the tests themselves, it proves impossible to evaluate the validity of the results obtained.

Sußmilch (1981) developed a vocabulary test for immigrant children in the 2nd and 3rd grades. The test aims at evaluating a subcomponent of a superordinate language competence, namely passive vocabulary, which is very important in learning a second language. Test construction began with a survey of available text books to develop a basic vocabulary list for immigrant children. For item construction nouns, verbs and adjectives with different frequencies were used in gap-filling and matching items accompanied by line drawings. The test combines 12 items for each part of speech and it takes 30 minutes.

This test was examined in empirical trials. In all, 426 non-German and 195 German pupils were tested in 29 classes from the Dusseldorf region. The test turned out to be rather easy for the immigrant children. The mean score was 27.1 (highest possible score 36) with a standard deviation of 7.0. For the immigrant children the test had a reliability of r_{tt} = .90 (alpha) and for the German pupils .82. The test is therefore a highly reliable measure of passive vocabulary, although the reliability sinks for the more advanced German group. The validity was assessed on the basis of teacher ratings of vocabulary knowledge and was calculated as r_{tc} = .41 for the immigrant group and .32 for the Germans. Therefore, the validity is not high and sinks further for the German group as the result of an obvious ceiling effect.

2.4 The C-Test in L2 testing

As I have shown the situation in L2 testing in the Federal Republic of Germany at the present time is most unsatisfactory. On the one hand there is an urgent need for economical, yet accurate tests, on the other hand there is no sign that any of the tests currently under development for use in L2 testing will be available for general use in a form fulfilling the normally accepted psychometric criteria in the near future. So, when I first heard the results Klein-Braley and Raatz were obtaining with the C-Test for their L1 and LF populations (cf. Klein-Braley and Raatz, this volume, and

Raatz and Klein-Braley, 1982), I decided to examine the performance of the C-Test for L2 groups. Raatz and Klein-Braley (1983) had already reported positive results in a pilot study, but there only 13 L2-subjects had been involved, so the results were suggestive but not concensive.

In the course of my investigations C-Tests were administered to a number of different immigrant groups. In my first investigation a total of 476 immigrant pupils, namely Greeks and Turks, and 197 German pupils took the tests. The test form used (D3A) had proved to be reliable and valid for German pupils (see Appendix). The C-Test was administered at the beginning of the school year 1982-83. The time limit was 40 minutes. For the non-German children the test proved rather difficult, and this was shown particularly by failure to finish the test. In the third grade, on average, 25 blanks were not attempted at the end of the test, in the fourth grade 10 on average. The mean difficulty for the immigrant children was P = 21 in the third and P = 40 in the fourth grade. Even so, all five test parts had high discrimination indices and the scores ranged from 0 to 83 for the immigrant children and from 1 to 98 for the Germans (out of 100 possible points). The test therefore differentiates well between individuals; nevertheless, it is obviously far too difficult for the 3rd grade pupils. Great differences emerged in the mean scores between different classes: the range lies between 6.3 and 37.6 in the third grade and from 16.3 to 65.0 in the fourth. Table 1 shows the reliability and validity coefficients (investigations 22 and 23).

Table 1

Reliability and validity coefficients
of a German C-Test with immigrant and German children

investigation	nationality	grade	N	r_{tt}	r_{tc}
22	immigrants	3	231	.86	.72
	Turks	3	92	.85	.61
	Greeks	3	107	.85	.74
	other	3	32	.83	-
	Germans	3	109	.92	.80
23	immigrants	4	245	.92	.69
	Turks	4	94	.87	.41
	Greeks	4	96	.92	.82
	other	4	55	.88	.73
	Germans	4	88	.88	.82

Values are given both for the total immigrant sample and for the individual subsamples (nationalities) per grade. The values of the German group permit comparisons to be made between L1 and L2 learners. As Table 1 shows the C-Test is highly reliable and valid. There are only slight differences between the values for the various

nationalities. The test is slightly less reliable for the immigrant children in the third grade and for the German children in the fourth grade. This reflects the fact that in the first case it was too difficult and in the second slightly too easy.

The validation criteria consisted of rating scales ranging from -5 to +5 for written and spoken German. The teacher was asked to use the entire scale to rate his/her pupils. The results were then averaged for each pupil and the validity coefficient for each class was calculated separately. For this reason the values in the column r_{tc} in Table 1 are the median values. Groups smaller than N = 8 were excluded so that no validity coefficients are given there. The other coefficients show that the C-Test is highly valid for all pupils with the exception of the Turkish pupils in the 4th grade. Here, however, the teachers involved in three classes had only known their pupils for less than eight weeks, and did not really feel competent to give an assessment of their pupils' proficiency in German. An additional factor with the Turkish pupils is the low variance in the test scores.

These investigations show that C-Tests can be used to assess language proficiency in German for foreign pupils and that the test is reliable and valid. The test can, in fact, be shortened to only 3 texts without serious losses in reliability. However, the three texts thus selected would be different for the different samples, and this would mean that test development procedures in the traditional fashion would be necessary.

In three further investigations two newly developed C-Tests were used. Test form DG is similar to D3A, but here the number of missing letters was indicated by providing a dash for each one. This, as Klein-Braley had shown in a pilot study, made the test easier. In investigations 24 and 25 this C-Test was administered to approximately half the Greek pupils involved in investigations 22 and 23. The average difficulty for this test was P = 52 in the third grade and P = 69 in the fourth. Table 2 shows the results of these investigations.

Table 2

Reliability and validity coefficients
of German C-Tests with immigrant children

investigation	nationality	grade	N	r_{tt}	r_{tc}
24	Greeks	3	52	.82	.58 - .74
25	Greeks	4	50	.85	.66 - .90
26	Turks	2	41	.94	.72 - .88

The reliability coefficients of r_{tt} = .82 (investigation 24) and .85 (investigation 25) are satisfactory, and again they can be

improved by discarding the less discriminating texts. The validity criterion was the same teacher judgements that had been obtained in the course of investigations 22 and 23, and Table 2 shows the highest and lowest validity coefficient obtained for the individual classes. Test form DG was an entirely new test using texts which had not been tried out before. It is therefore interesting to compare the results obtained for the two C-Tests administered four months before to the same subjects. The correlations for the two sets of scores are $r = .81$ for the third grade and $r = .90$ for the fourth grade. The two C-Tests are therefore parallel tests although there was a time lapse of four months between the administrations, item analysis had not been performed, the deletion technique had been modified and entirely different texts were involved in the two investigations!

For investigation 26 the attempt was made to reduce text difficulty still further. The texts were selected from readers for the first grade or from L2 text books for the first year of teaching. The texts were shorter and had only ten deletions which affected every third word. Again a dash indicated each missing letter. Test form DT contained 6 texts of this type and was administered to 41 Turkish pupils towards the end of the second grade. They were given 60 minutes to complete the test, the first pupil finished in 13 minutes and, on average, they needed 40 minutes. Table 2 shows the results of this investigation (26).

Text difficulty ranges from P = 26 to P = 55, the mean difficulty is P = 40. Test reliability is $r_{tt} = .94$, the highest reliability coefficient so far obtained for a C-Test. Even when the test is shortened to only 3 texts (30 deletions) the reliability drops slightly to $r'_{tt} = .91$. Again teacher ratings were used for validation, and the values for the three classes range from $r_{tc} = .72$ to .88, which shows very high agreement between teacher judgements and C-Test scores.

The investigations into the performance of C-Test for L2 learners in Germany have produced highly satisfactory results. Three different test versions were used and each was shown to be reliable and valid. Once again it has proved possible to develop highly reliable and valid tests with no preliminary test development procedures. The C-Test results show high agreement with teacher ratings. In view of the extreme heterogenity of the L2 groups in German schools it would seem desirable to produce test versions for specific target groups. In particular, extremely easy tests are needed for the early stages of L2 learning, but as has been shown in investigation 26, modification of the C-principle enables the construction of suitable tests for these subjects.

3 Proficiency testing with L2 groups in Germany

The approaches currently available in the Federal Republic of Germany differ not only in their linguistic basis and their pedagogical aims, but also in the degree of standardisation and empirical quality control desired. One approach demands a curriculum-integrated diagnosis of language development aiming at a highly differentiated

description of the individual level of proficiency in L2 in order to determine appropriate therapeutic procedures. A second approach demands the development of standardised language tests which are objective, reliable and valid. As the investigations reported here show the C-Test fulfils all the requirements of this second approach. The C-Test can be used by teachers in normal classroom test procedures, but it can also be used as an aid in selection, classification and placement decisions.

The lack of suitable standardised tests for L2 groups in Germany has provoked the development of informal test procedures in this area by practitioners working in the field. Unfortunately, the majority of these procedures have not been subsequently examined to determine their test qualities. Furthermore, tests for which no details are available regarding item characteristics, administration and scoring objectivity, reliability and validity must be regarded as dubious procedures. The majority of these informal procedures have the additional disadvantage that they must be administered individually and can take up to two hours.

Standardised language tests such as the C-Test on the other hand have the immense advantage that they are economical. The C-Test can be shortened so that only 20 to 30 minutes are needed for test administration, thereby facilitating its use even in the second grade, yet test reliability remains satisfactory. Placement decisions could be made very quickly using the C-Test, and since one of the main problems in teaching L2 learners is the setting up of homogeneous groups for more effective language instruction, the C-Test could play a very important role in improving the educational chances for immigrant children.

Appendix

Examples of C-Tests

1 C-Test D3A (text 3)

Ein junger Hase, der zum erstenmal in die Sonne kam, sah plötzlich seinen Schatten neben sich. Hals üb_____ Kopf ran_____ der Ha_____ davon, do_____ das schw_____ Tier m_____ den lan_____ Hornern ran_____ ebenso sch_____ neben i_____ her. D_____ Hase sch_____ Haken na_____ rechts u_____ links u_____ jagte i_____ den dun_____ Wald zur_____. Da end_____ war se_____ Verfolger verschwunden. Atemlos und erschöpft hielt der Hase an.

2 C-Test DG (text 1)

Das kleine Mädchen wollte so gern einmal mit dem Vater wippen. Es

hat sich auf die eine Seite von der Wi _ _ _ gesetzt, u _ _ der Va _ _ _ hat si _ _ auf d _ _ andere Se _ _ _ von d _ _ Wippe ges _ _ _ _. Da w _ _ das kle _ _ _ Madchen ga _ _ hoch ob _ _ , und d _ _ Vater w _ _ ganz un _ _ _ . "Los! L _ _ !" hat d _ _ kleine Mad _ _ _ _ gerufen. Ab _ _ die Wi _ _ _ konnte nicht wippen. Das kleine Madchen war doch so leicht, und der Vater war so schwer.

3 C-Test DT (text 6)

Im Garten steht ein großer Kirschbaum. Die Kirschen sind reif. Vi _ _ _ Spatzen sitzen i _ den Zweigen. S _ _ picken die Kirs _ _ _ _ an und mac _ _ _ ein Geschrei. D _ _ Nachbar kommt a _ _ dem Haus. E _ schimpft: Diese fre _ _ _ _ Spatzen! Ich m _ _ die Kirschen heute noch pflucken, sonst sind sie alla angepickt.

Bibliography

Akademie für Lehrerfortbildung Dillingen (Hrsg). No date. Deutschunterricht für Kinder ausländischer Arbeitnehmer. Ein Fortbildungsmodell. Donauwörth: Auer.

Biehl, J. 1982. Sprachstand und Sprachlernbedingungen türkischer Schüler. Unveröffentlichtes Manuskript. Duisburg: Universität Duisburg Gesamthochschule.

Damanakis, M. 1978. Sozialisationsprobleme der griechischen Gastarbeiterkinder in den Grund- und Hauptschulen des Bundeslandes Nordrhein-Westfalen. Kastellaun: Henn.

Fliegner, J., and I. Gogolin in Zusammenarbeit mit R. Urbanek. 1982. Sprachstandsmessung bei Schulanfängern. Projektberichte des Pädagogischen Instituts der Landeshauptstadt Düsseldorf, Heft 5. Neufassung nach Überarbeitung. Düsseldorf: Pädagogisches Institut.

Hegele, I. 1981. "Deutsch als Zweitsprache: Sprachstandsdiagnose und Sprachförderung von Ausländerkindern". Deutsch lernen 6: 41-56 (Heft 4).

Ihssen, W.B. 1980. "Probleme der Sprachentwicklungsdiagnose bei Ausländerkindern. Hinweise auf Voraussetzungen zuverlässiger Ergebnisse". Praxis Deutsch, Sonderheft 1980 - Deutsch als Zweitsprache: 40-42.

Luchtenberg, S., Neumann, H.-J., and M. Wespel. 1982. "Informeller Test zur Sprachstandsdiagnose bei ausländischen Kindern und

Jugendlichen". In Turkische Kinder in unseren Schulen. Eine padagogische Herausforderung. Mit Beiträgen von U. Coburn-Staege ... Stuttgart: Klett. 134-176.

Mitzschke, M. 1978. "Probleme der Eingruppierung auslandischer Jugendlicher im Deutschunterricht. Ein Praxisbericht". Diskussion Deutsch 9: 77-81.

Portz, R., and C. Pfaff. 1981. SES - Soziolinguistisches Erhebungsinstrument zur Sprachentwicklung. Berlin: Padagogisches Zentrum Berlin.

Raatz, U., and C. Klein-Braley. 1982. "The C-Test: a modification of the cloze procedure". In Culhane, T , Klein-Braley, C., and D.K. Stevenson (eds). 1982. Practice and problems in language testing 4. University of Essex. Occasional papers 26. Colchester: University of Essex Press. 113-138.

- 1983. "Ein neuer Ansatz zur Messung der Sprachleistung. Der C-Test: Theorie und Praxis". In Horn, R., Ingenkamp, K., and R.S. Jager (Hrsg). 1983. Tests und Trends 3. Jahrbuch der Padagogischen Diagnostik. Weinheim: Beltz. 107-138.

Scherzinger, A., and G. Scherzinger. 1981. "Sprachstand-Diagnose fur die schulische Forderung von Auslanderkindern". Auslanderkinder in Schule und Kindergarten 2: 18-22 (Heft 1).

Steinmuller, U. 1983. "Sprachstandserhebungen in Berlin. Bericht uber zwei Projekte". In Regionale Arbeitsstellen zur Forderung auslandischer Kinder und Jugendlicher (Hrsg). 1983. Sprachstandserhebung bei auslandischen Kindern und Jugendlichen. Tagungsbericht vom 26.11.1982. Essen (erscheint in Kürze).

Süßmilch, E. 1981. Der Sprachstand ausländischer Grundschuler. Entwicklung eines Wortschatztests und seine empirische Prufung. Unveroffentlichte Diplomarbeit. Düsseldorf: Universität Düsseldorf.

'CLASSROOM LANGUAGE' - A NEGLECTED AREA IN FOREIGN LANGUAGE TEACHING AND TESTING

Bernd Voss (University of Bielefeld)

1 Introduction

This paper is an attempt to give an account of some aspects of the background to a larger project concerned with the development of teaching materials in the field of 'classroom language' for German teachers of English (1).

By 'classroom language' we mean the kind of language used by teachers and pupils to conduct classroom business, e.g. to set up groups for group work, to distribute material, to organise activities, to tell a pupil off, to focus attention onto a particular teaching point, to ask for further clarification or for a repetition, to bid for the floor, to express joy or regret over something that has just happened in the classroom etc.. All of these are real communicative acts between teacher and pupil, pupil and teacher, or amongst the pupils themselves, within the classroom setting. Conducting this kind of interaction (as far as possible) in the target language opens up a directly relevant practice field in using the foreign language for real communicative purposes. However, language teachers do not always make (appropriate?, full?, any?) use of this only genuine situation that institutionalised language teaching provides.

A major reason for this neglect appears to be that language teachers themselves tend to have difficulties with this domain of language, and we shall argue in this paper that there are mainly three closely interrelated factors responsible for this - (i) deficiencies in the set-up of the language education of intending language teachers (certainly in Germany); (ii) lack of suitable teaching materials for this language domain; and (iii) uncertainty as to what 'classroom language' actually entails.

In what follows we shall take up each of these three problems in turn inan attempt to clarify the field before providing a brief sketch of the nature of the proposed teaching and testing materials and some general conclusions.

2 Language Needs and Language Education of Language Teachers

An attempt to specify the language needs of language teachers can perhaps best start off from the observation that teaching a foreign language is rather different from teaching any other school subject in that the foreign language is not only the goal but also the medium of instruction. In subjects such as biology, mathematics, or sports, where the objectives may be knowledge of biological facts, understanding of mathematical processes or the development

of ergonomic movements, the medium of instruction is the mother tongue. This is typically different in the modern language classroom where the foreign language is to replace the mother tongue as the medium of instruction as soon as practicable, and that could mean right from the start.

While this may not be universally accepted as a description of the facts at present, there is little doubt about its desirability in principle. Few people would disagree with the view that the aim of teaching a modern language has not been reached when learners are able to talk about the language and its cultural background in their mother tongues, but that it is necessary for them to develop the ability to interact meaningfully in it. Where is the learner to learn this if not in the foreign language classroom? But how is this conceivable unless the interaction in the classroom is done as far as possible in the foreign language itself?

There is no doubt that conducting lessons entirely in the target language makes particularly high demands on the teacher's active command of the foreign language, and it may be instructive to look more closely into the language needs of language teachers in general, before considering the question of how and whether these needs are being met within the context of teacher education.

The language needs of the language teacher can be conveniently thought of as having a general aspect, a study-related aspect and a classroom-related aspect.

Since it is not central to our argumentation, it may suffice here to characterise the general aspect of the language command hopefully possessed by a language teacher in such general terms as the ability to cope with everyday communicative needs in the country of the target language, or to explain aspects of one's own country to a foreigner. The need for this aspect of language command is obvious in that this is what teachers are usually supposed to teach their language learners in the foreign language classroom.

The study-related aspect of a teacher's language command refers to specialised language domains such as the language of literary analysis, of linguistic description, of methodological discussion, of the presentation of cultural and sociological fields. Language teachers need this aspect of their language command in order to be able to inform themselves about their own fields of study in the target language, but also to be able to introduce the more advanced ones of their own learners to the basics of the disciplines concerned. This need is perhaps most obvious for teachers teaching advanced learners about to enter university, i.e. at a level where highly demanding topics call for appropriate linguistic skills on the part of the teacher.

The classroom-related aspect of a language teacher's language command is the component which we are particularly concerned with here. It refers to the kind of language needed to conduct lessons in the target language, i.e. to cope with the classroom situation

in the foreign language. Why teachers need this aspect of their language command has already been suggested above: there are many good reasons why foreign language lessons should be conducted as far as possible in the target language. What this entails will be set out in more detail below (part 4), so that we can now turn to the question of how these various language needs are met within the context of teacher education.

Teacher education in Germany usually starts with a three or four year course at a university or college (1st stage). This is then followed by a probationary period of between 18 months and two years, in which the trainee teacher receives practical training in the teaching of his or her subjects (2nd stage). The examination taken at the end of this stage establishes a person as a fully qualified teacher. At a later stage, there is in-service training available for those who are lucky enough to be near where such courses might be offered and sufficiently interested in attending (3rd stage).

Looking at these three stages of teacher education in reverse order makes plain how much the language education of future language teachers is a chance product rather than the result of careful planning.

In-service courses - as far as they exist at all - usually exclude training in the target language, presumably because the addressees are qualified teachers who by definition must possess an acceptable level of competence in the foreign language, which obviates the need for further activities in the field. In other words, it would come near to breaking a taboo to offer language courses of any kind at this level because such a step would recognise officially that deficiencies exist.

A similar reluctance to accept explicit language training as part of the instructional programme can be observed during the second stage of teacher-training. By passing their university examination, trainee teachers have demonstrated their possession of at least passable levels of language competence in the foreign language, which could not possibly be called into question by offering more language instruction at this stage. To be fair, there are also many other important demands competing for time at this stage, but in spite of not infrequent complaints about the limited language command of some future language teachers (cf.e.g. Ostberg 1982), not much is explicitly being done about it.

Even at the first stage of teacher training it should not be forgotten that it is only comparatively recently that the language teaching aspect has been taken somewhat more seriously. Especially where languages such as English or French are concerned with their long standing and strong presence in German schools (often between seven and nine years of instruction before entering university), tradition is still strong in encouraging the belief that language teaching is not really a proper task for a university - apart perhaps from translation - , since students can theoretically be expec-

ted to have a reasonable command of the language when they arrive, guaranteed by the final school-leaving examination, which effectively defines the end of language learning needs. Where does the buck stop?

Turning now, against this outline of the language education of future language teachers, to the question of how their language needs are met within the context of their training, we are confronted with the following picture.

The general component - perhaps precisely because of its inherent vagueness - is the least problematic one in this context in that it has come to be an accepted part of the training at the first stage. While not all teachers eventually turn out to be proficient in this field, there seems to be general agreement nowadays that improvement is necessary, and steps are being taken to remedy the situation by encouraging periods of residence abroad and by providing appropriate language courses.

There is less readiness to cater for the study-related language needs of future language teachers. The possible step of teaching all or at least most subjects usually making up a teacher's study course in the target language is hardly ever taken in German universities and colleges. Still, students do become at least somewhat acquainted with this aspect of the language command expected of future teachers, in the context of their studies.

It is the classroom-related aspect of their language command that future language teachers get least preparation for. As pointed out above, at the second and third stages of teacher education, overt language training is practically nonexistent. At best, trainee teachers are referred to one of the numerous little booklets with more or less adequate collections of classroom phrases (cf. part three below for more details), which do little to improve the situation. A significant effect can only be expected from explicit teaching at the first stage, where overt language teaching is considered acceptable. However, in spite of the considerable progress that sociolinguistic research into aspects of language use in the classroom has made in recent years (cf. Sinclair and Coulthard 1975, Edwards and Furlong 1978, Mehan 1979, Sinclair and Brazil 1982, Lörscher 1983), the growing awareness of the role of language in classroom teaching seems to have had little influence to date on the language training of future teachers of foreign languages. It is extremely rare to find institutions which expressly include classroom-specific language needs in their language teaching programmes, nor are there usually any steps taken to specifically screen-test future teachers as to their proficiency, or lack of it, in this highly relevant field of language use.

It seems that classroom language has not yet emerged in the professional consciousness as an important aspect of the language command of (future) language teachers. Two fallacies contribute to the frequent failure to perceive the problem. One is implicit in

the notion that a good general command of the language will carry the language teacher through most of the exigencies of his professional life. While nobody would quarrel with demanding a high general level of attainment, unfortunately not all language teachers can be said to have reached this stage in the language they are teaching. In addition, however, the teaching situation is characterised by interactional patterns rarely found outside the classroom, so that the non-native teacher will have little experience with classroom talk no matter how good his general command of the language might be. The second fallacy is implicit in the curious notion that certain aspects of the verbal interaction in the classroom are not legitimately part of the lesson 'proper' and should therefore be conducted in the mother tongue of the learners rather than in the foreign language. We shall comment on this more fully below (part 4).

We have tried to show so far that classroom language has not yet found its proper place in the context of teacher education. The result of this neglect is that important aspects of the verbal interaction in the foreign language classroom are not at all, not sufficiently or not adequately covered in the target language, with unfortunate effects on the language learning opportunities of the pupils or students involved.

Observations of lessons taught by university students on teaching practice and even by fully qualified (non-native) teachers show that deficits in the classroom-related component of the teacher's language command tend to make themselves felt in one or the other of the following ways.

The teacher may use the mother tongue for a large proportion of the interaction in the classroom, thereby depriving his learners of considerable opportunities to use the foreign language meaningfully. This is primarily a quantitative problem, but we may also expect this to induce in the learner the - perhaps unconscious - notion that when it comes to real communicative needs the target language is unsuitable, so that it would also be a qualitative one.

On the other hand, the teacher may take over the kind of language he himself experienced as a learner, with all its mistakes and deficiencies, and thus establish and perpetuate a classroom dialect at worst comprehensible only within that particular classroom, without ascertaining its value in the real world.

Or the teacher may use a tightly controlled repertoire of stereotyped formulae which in their rigidity (blunt imperatives prevail) are not only inappropriate but also dangerous as a model for the learner inasmuch as the latter may be tempted - or forced through lack of experience with alternatives - to use similar expressions himself (e.g. 'Speak louder!'), for instance in a conversation with his peers or even superiors, without being aware of the social implications of the phrase used.

3 Classroom Phrase-Books

If the first of the factors contributing to the feelings of unfamiliarity and unease towards the domain of classroom language is connected with deficiencies in the set-up of the language education of future language teachers in Germany, there is no doubt that lack of suitable learning and teaching materials for this area is the second.

As we have noted above, it is not unusual, during the second stage of teacher education, for the teacher trainer or perhaps an older colleague to draw the trainee teacher's attention to one or another of the numerous booklets containing lists of classroom phrases arranged under various headings and intended for the use of the practising teacher (2). Such a move is often triggered off by some language mishap that might have occurred in the trainee's lesson, and it is equally often carried out in a somewhat embarrassed and apologetic atmosphere: the commendable attempt to spare a young colleague's feelings appears mingled with the shameful recognition of the inadequacy of the help offered. The booklets are not really considered 'respectable', and are passed on as covert rather than overt information.

The idea of the phrase-books is that they are to be used for reference purposes if, during the preparation of the foreign language lesson (or also after the event), the teacher realises that there is something which he intended to do and which he cannot (or could not) confidently do in the foreign language. In other words, their usefulness depends entirely on the realisation that something might go, or has gone, wrong. Considering the characteristic ways in which problems with the domain of classroom language manifest themselves (cf. part 2 above) it comes as no surprise to find that the phrase-book approach does not seem to make much impact. Although the dictionary format (e.g. Kissling 2/1981) is the exception rather than the rule, it is the logical consequence of the reference purpose the books are primarily meant to serve. However, dictionaries by themselves are not very suitable learning and teaching materials in this field, nor in any other.

The other and perhaps more serious reason for the understandable misgivings aroused by the phrase-books is connected with their contents. While a more detailed analysis than is possible here would reveal the various merits and demerits of the individual phrase-books concerned, in general they appear narrow in scope, often unreliable in their information, and unsystematic in their view of the nature of what is presented as classroom language. The result is that many questions a user might have, are in fact left unanswered.

4 Classroom Interaction and Classroom Language

The third factor contributing to the problem of classroom language in foreign language teaching - and the difficulties teachers

themselves may have with this language domain - is connected with a more general uncertainty as to what the notion of 'classroom language' entails.

Language has an extremely important role to play in any educational setting, and perhaps even more so in foreign language lessons. Looking at these lessons in their entirety and at the kinds of verbal interaction going on within them immediately draws attention to the fact that the language of the classroom is by no means all teacher talk, but includes pupil-teacher talk and, of course, also pupil-pupil talk. In addition, it is important to realise that a considerable proportion of the interaction in the classroom is not directly concerned with the topic of the lesson at all, but with preparatory moves, organisatory aspects, social behaviour, or similar fields of reference. In other words, classroom language covers a much wider field than just one aspect of a teacher's language command.

Existing implicit (as e.g. in classroom phrase-books) and explicit (cf. theoretical literature) categorisations of the field can be systematised into five different - and no doubt complementary - approaches which, taken together, provide a comprehensive mapping of the classroom language domain and can be captured in the following five questions,

1 At what stage in the lesson does the interaction take place?

2 What is the interaction about?

3 What is the function of the interaction in the context of the lesson?

4 What is the communicative purpose of the interaction?

5 By which linguistic means is the interaction implemented?

We shall now discuss each of these five questions in turn in an attempt to provide a comprehensive outline of the field, as well as an indication of where the problematic areas of the concept of classroom language lie.

4.1. At what stage?

The first question draws attention to the sequential structure of a lesson,the stages of which are often separated from each other by explicit boundary signals, such as "Now let's close our books and turn to the board" or similar indicators. Most classroom phrase-books follow this approach in the presentation of the greater part of their material. Stages suggested by this approach include the following:

1 opening (making the pupils come in, greetings; attendance, lateness; opening of the actual proceedings)

2 transition (from one stage to the next)

3 homework (setting, checking)

4 working with course books (texts, exercises)

5 working with/at the board

6 working with technical aids (OHP, tape recorder, cassette recorder, language laboratory, video; maps, charts, flashcards etc.)

7 working with worksheets (producing them, distributing them in class, using them during the lesson)

8 songs and games

9 group work and role play

10 discussing topics, organisation, behaviour

11 discussing extra-curricular activities (excursions, day trips)

12 topic-related work (usually based on texts)

13 language-related work (pronunciation, orthography, syntax, lexis, pragmatics; comprehension; conversation skills etc.)

14 tests and examinations (announcing, setting up, correcting, discussing)

15 announcements (speech day, assembly; changes in time-table, absent teachers, temporary replacements)

16 special events (festivities, holidays, prizes, birthdays etc.)

17 closing (end of work, clearing up, valediction, dismissal)

While it is no doubt possible to define these stages differently, an attempt has been made in the above list to be as comprehensive as possible and to present exhaustively all characteristic stages of a lesson. What is perhaps more interesting than the question of whether they do indeed cover all conceivable cases is the fact that they mirror different levels at which decisions about the next stage in a lesson tend to be taken. While some are definitely bound to a particular place in a lesson (1, 17), others are free as to when they occur and are characterised by the medium used (4 to 7), or the type of activity envisaged (8 to 11), or the main purpose pursued (12, 13), or the function fulfilled (2, 3, 14, 15, 16). These differences of level may be theoretically unsatisfactory, but are a natural result of the fact that different kinds of features can serve as the dominating aspect of a particular stage of a lesson.

4.2 What about?

The second approach to classroom language starts off from an analysis of what the interaction in the classroom refers to, i.e. what it is that teachers and pupils in the classroom in fact talk about. It shows that it is necessary to distinguish between several areas of reference, within the context of the verbal interaction in a lesson.

Priesemann (1971) distinguishes between subject-related and communication-related talk in the classroom setting, emphasising that a considerable proportion of the verbal interaction is needed to establish common ground and to prepare for the understanding of the topic, and that without these moves the learner has no access to the subject-related part of the interaction. More specifically concerned with foreign language teaching is Black and Butzkamm's (1977) distinction between language-related and content-related aspects of classroom communication. Language-related communication is characterised as monodimensional: the main concern here is whether language has been used formally correctly. Content-related communication, on the other hand, is more concerned with the content of an utterance as an expression of a communicative aim. The examples and further explanations narrow this down to refer specifically to the real-life situation of the learner, even to the exclusion of e.g. the lesson topic. Fanselow (1977) is more precise in this respect in distinguishing between language/life/procedure/subject matter as areas of reference of classroom talk. Language refers to a wider field than Black and Butzkamm's language-related communication, because Fanselow sees metacommunication as part of this category, whereas in Black and Butzkamm's analysis this is expressly excluded. Procedure highlights those areas of classroom talk that are concerned with procedural questions. The separation of life and subject matter into two categories makes it possible to keep the real-life situation of the learner apart from the content of the lesson, as two different fields of reference of classroom discourse. Chaudron's (1977) analysis of the correction behaviour of foreign language teachers shows that the corrections refer either to language points (linguistic), or to the actual classroom situation (classroom interaction and discourse) or to the lesson content (subject matter). His second category then covers what Fanselow keeps separate as life and procedure. From a somewhat different point of view, Boueke, Henrici and Schülein (1983) describe classroom interaction as either content-referenced, process-referenced or group-referenced. Content-referenced interaction focusses on the lesson topic, process-referenced interaction aims at establishing acceptable procedures according to which the topic is to be handled in class, and the group-referenced parts of the interaction refer to the development of behavioural norms within the group of learners. The aspect of behaviour as something that is potentially being talked about in the classroom is also singled out specially in Sinclair and Brazil's (1982) analysis, where distinctions are made between the topic (content), the organisation of working procedures (organisation), the behaviour of the learners (discipline), and - although perhaps not very distinctly put - classroom dis-

course (discourse) itself, as areas which teachers and pupils talk about in the classroom setting.

An attempt to systematise the results of the above studies leads to the following six areas of reference of classroom discourse:

1 topic. This term covers the lesson contents, or topic, as it is for instance suggested by the text or textbook used, i.e. the (non-actual, imaginary) situations, problems, facts presented for classroom treatment. Examples:

Who was Mary Stuart? When does Peter get up every morning? What is the main point of the first part of the text?

2 reality. This term covers references to the real-life situation in which the learners find themselves (except areas five and six below). We would argue that this area is of primary immediacy to them, as opposed to the secondary degree of immediacy represented by the first field above. Examples:

It's nice to see you back, Peter (after an absence). How old are you, Mary? Could I borrow your rubber? May I go to the toilet? Please, sir, Henry's cut his finger.

3 language. The third area covers explicit references to formal aspects of the language being taught, either as the focus of the teaching at that stage, or in the context of a correction. Examples:

Remember, committee is spelt with two e's at the end. What is the past tense form of teach again, Barbara? What is the meaning of dowdy?

4 discourse. This term is meant to capture those parts of the verbal interaction in the classroom explicitly concerned with the establishment and maintenance of communication. Examples:

I couldn't hear you - could you say that again, please? Is that what you mean? I'm not sure whether I've made myself understood, but what I mean is... Could you explain your last point again?

5 organisation. This area covers utterances directly concerned with organisatory moves. Examples:

Could you do the next sentence? I've finished - what am I supposed to do now? May I work with Colin on this? Could you share your book with Linda today - we're one copy short. I haven't heard my turn yet. Could you split up into groups of four now - we want each group to work on a particular aspect of the story.

6 discipline. This final area refers to the parts of classroom talk concerned with the maintenance of order. This is usually initiated by the teacher, but can also be sparked off by an appeal from a pupil. Examples:

Could you pay attention for once. No prompting please: give him a chance to do it himself. Please, sir, Bob keeps pulling my hair.

A recent informal survey among students studying to become teachers of English (N = 20) revealed that they thought the following classroom activities to be the three most important ones for language teachers to be able to do in the foreign language: (1) explain the meaning of new words, (2) explain a grammatical construction, (3) correct oral contributions. Lowest ratings were given to: explanations about classroom procedures; setting up of technical aid; checking attendance and doing other administrative duties. While areas 3 and, of course, 1 tend to be accepted as fields of reference to be covered in the foreign language, in foreign language teaching, many of the other areas are often not considered to be part of the lesson 'proper' and are therefore dealt with in the mother tongue.

However, it can hardly be emphasised enough that including these other areas as widely as possible in the foreign language part of classroom discourse is a decisive aspect of the credibility of teaching a foreign language at all: the more these fields are excluded from communication in the foreign language, the more the learner is inevitably left with the impression that real communicative purposes can only be served in the mother tongue, and that the foreign language is only suitable for the verbalisation of the necessarily fictitious worlds of textbooks and other teaching materials.

4.3 What for in the lesson context?

The third approach to classroom language takes its starting point from the question of what a particular part of the interaction means in the context of a lesson. In a rather detailed analysis, Sinclair and Brazil (1982) distinguish between four basic functions, initiation as the starting point of an interactional sequence, response as the reaction to an initiation, follow-up as the reaction to a response, and frame and focus, signalling a transition from one stage to another. A sequence of the first three functions is called teaching exchange, the last two are grouped together as boundary exchange. Bidding, i.e. asking for permission to speak, can be considered a special aspect of the response function; complex initiations include nominations, cues, prompts, clues, and of course reformulations as re-initiations.

It seems, then, that it is useful to distinguish between (1) initiation, (2) response, (3) follow-up and (4) boundary as the four basic functions of classroom discourse within the context of a lesson. Classroom conventions suggest that all except the second

are dominated by the teacher. This is not unproblematic, in particular in foreign language teaching, because although it enhances the (desirable) role of the teacher as a language model, a strong dominance and the resulting social markedness of the language used may imply unsuitability of the teacher's language for the purposes of the learner. As far as possible, care must therefore be taken (i) to involve the learners also in functions other than the second, (ii) to use language which is 'reversible', i.e. which can be used reciprocally, as between equals, rather than language which is clearly marked for hierarchial (social) implications (as e.g. commands, in contrast to requests, etc.).

4.4 What for in the context of communication?

Asking for the purpose of a part of the interaction in the classroom is not only a way of aiming at sequential structural elements of a lesson (as above), but can also serve to elicit general communicative functions. Thus, Bellack et al. (1966) distinguish between four moves within classroom communication, structuring, soliciting, responding and reacting. A comparison with the approaches described above shows, however, that structuring captures the reference to a particular field of discourse, viz.organisation (see 4.2 above), whereas the other three correspond to a teaching exchange as described in the previous section. More recent is Sinclair and Brazil's (1982) distinction between four communicative functions of teacher talk, telling, controlling, stimulating and rewarding. The second and third of these correspond to Bellack's soliciting move and imply a physical response in the first case and a verbal one in the second. Rewarding apparently covers what is termed reacting by Bellack. Telling is specifically concerned with the communicative function of informing. All others are focussed on structural lesson elements, which is perhaps not quite so obvious since Sinclair and Brazil's description - with its emphasis on the language of the teacher - does not include the reaction of the learners in the analysis.

Looking at the possible communicative functions of language within a lesson in perspective, we propose the following categorisation in an attempt to do justice to the typical range of aspects involved:

1 inform. This function is not only fulfilled by the teacher passing on a piece of information (e.g. 'Washington is the capital of the United States'), but also by the pupil who answers a (factual) question.

2 evaluate. It will usually be the teacher who deals out praise or blame. However, this function would also be performed more neutrally by pupils and teacher alike when e.g. expressing personal feelings, making personal comments ('what a dirty trick that was') or making explicit value judgments ('I don't like this story very much because ...').

3 instruct. What we mean by this term is the function of someone (usually the teacher) telling someone else to do something, as opposed to

4 elicit, by which we mean the function of asking someone to say something, i.e. to give a verbal response, as is customary with most teacher questions.

5 maintain social climate. With this category we try to capture those parts of the verbal interaction in the classroom that are not obviously concerned with any of the above and primarily serve social functions, such as greetings, valedictions, encouragement ('You're nearly there - just carry on'), social rituals in connection with e.g. a birthday ('Many happy returns'), an excuse ('That's all right'), a sneeze ('Bless you'), a piece of bad news ('I'm sorry to hear that').

6 maintain communication. This category is intended to cover utterances that influence and steer the communication processes in the classroom, such as digression signals ('incidentally ...'), appeals for agreement ('we are all agreed, aren't we, that ...'), handicap signals ('I haven't perhaps understood it all, but has this to do with ...'), appeals for (linguistic) help ('I forgot the English for it again, but what I mean is ...'), or similar.

Obviously, these six communicative functions are by no means classroom-specific as such, although their relative frequency here will differ from that found e.g. in private communication. However, we wish to emphasise that also categories 5 and 6 have their legitimate place in the classroom context: if the aim of foreign language teaching is the eventual ability of learners to interact naturally in the target language, it is difficult to see how this could be achieved if these two functions were to be systematically excluded from foreign language classroom discourse.

4.5 How?

The last question approaches the verbal interaction in the foreign language classroom from the point of view of the linguistic forms needed to carry the various functions described above. It is well known that there is no one-to-one relationship between form and function, and the following forms can serve a variety of - sometimes rather classroom-specific - functions:

1 declaratives, e.g. The blackboard hasn't been cleaned. You are too noisy. I'd pick that up if I were you.

2 interrogatives, e.g. Could you do the next sentence? Why don't you keep quiet.

3 imperatives, e.g. Try again, Peter. Don't give up.

4 interjections, e.g. My goodness, what a noise. Right. Crumbs.

Table 1

Aspects of verbal interaction in the (foreign language) classroom

1. At what stage in the lesson?		2. What about?	3. What for? (lesson)	4. What for? (communication)	5. How?
1 opening		1 topic	1 initiation	1 inform	1 declaratives
2 transition		2 reality	2 response	2 evaluate	2 interrogatives
3 homework			3 follow-up	3 instruct	3 imperatives
		3 language		4 elicit	4 interjections
4 course book	work with/ at	4 discourse	4 boundary		
5 blackboard				5 maintain social climate	5 set phrases
6 technical aid		5 organisation			6 signals
7 worksheet		6 discipline		6 maintain communidation	
8 songs, games					
9 group work, role play					
10 discussing: topic organisation, behaviour					
11 discussing: extra-curricular activities					
12 topic - related work					
13 language - related work					
14 exams, tests					
15 announcements					
16 special events					
17 closing					

5 set phrases, e.g. Happy birthday to you. You're welcome.

6 signals, e.g. hesitations, intentional changes in pitch ('quotation'), but also linguistic signals (below the sentence level) such as: however, on the whole, as it were.

5 Teaching and Testing Materials

On the background of the preceding survey it is now possible to sketch out in broad terms the basic requirements suitable teaching and testing materials need to fulfil.

First, as our survey has shown, classroom language is a wider concept than is often realised, and this needs to be reflected in the materials offered for this language domain.

Second, the materials need to be performance-oriented in that they need to teach not only the knowledge of the language used in the classroom, but also the ability to use it.

This aim would assign a very important place to criterion-referenced tests as a check on the knowledge aspect, and a prerequisite for the performance-oriented aspect, of a language course in this field.

Finally, it would seem important to develop a comprehensive screen-test for this language domain in order to make sure that (future, or even present) language teachers who are themselves uncertain of their language command in this field, can be given specific help.

6 Conclusions

We have argued in this paper that classroom language still has to come into its own as an important aspect of foreign language learning and teaching. We hope to have been able to show that it can make an important contribution to meaningful foreign language instruction.

Providing for overt training in this field implies not only the provision of adequate information and practice, but also, in many cases, an attempt to change attitudes, e.g. as far as the appropriateness of certain areas of reference for foreign language classroom discourse is concerned. Teaching styles and methods would no doubt be rather directly affected, too.

Although there are obviously vast areas of overlap between what we have called the general and the classroom-related aspects of a language teacher's language command, we believe that classroom language provides a less arbitrary, and therefore possibly more helpful, framework for language learning activities within a teacher's training course, than whatever general language compe-

tence is taken to mean: any time redirected from general language courses to a specific one on classroom language would not jeopardise the general standard, but would help to raise it by providing language work within an obviously relevant context.

Recent course materials (Hughes 1981, Willis 1981) show that important progress has been made in the conceptualisation of what classroom language implies and in the presentation of it for learning and teaching purposes. It is to be hoped that the project, the background of which has been presented in this paper, will be a further step in this direction.

7 Notes

1 supported by grant No 8546 of the University of Bielefeld.

2 e.g. Müller and Zellmer (1952), Windmann (1959), Meldau and Pearcy (1961), Weidner (1968), Heuer and Parry (1971), Ebbinghausen (1977), Kißling (2/1981), Gressmann and Rich (1982).

8 Bibliography

Bellack, A.A., Kliebard, H.M., Hyman, R.T., and F.L. Smith (eds). 1966. The language of the classroom. New York: Teachers College Press.

Black, C., and W. Butzkamm. 1977. "Sprachbezogene und inhaltsbezogene Kommunikation im Englischunterricht". Praxis 24: 115-124.

Boueke, D., Henrici, G., and F. Schülein. 1983. "Handlungsfunktionen und Diskursgliederung in selbständig arbeitenden studentischen Lerngruppen". In Ehlich, K., and J. Rehbein (eds). 1983. Kommunikation in Schule und Hochschule. Tübingen: Narr. 137-155.

Chaudron, C. 1977. "A descriptive model of discourse in the corrective treatment of learners' errors". Language Learning 27: 29-46.

Ebbinghausen, K. 1978. Classroom communication. Paderborn: Schöningh.

Edwards, T., and V.J. Furlong. 1978. The language of teaching. London: Heinemann.

Fanselow, J. 1977. "Beyond Rashomon: conceptualising and describing the teaching act". TESOL Quarterly 11,1: 17-39.

Gressmann, L., and A. Rich. 1982. Classroom language. Bielefeld: Cornelsen.

Heuer, H., and P. Parry. 1971. Hands up! Unterrichtssprache deutsch-englisch. Dortmund: Cruwell.

Hughes, G.S. 1981. A handbook of classroom English. London: Oxford University Press.

Kissling, H. 2/1981. Lexikon der englischen Unterrichtssprache. Heidelberg: Quelle & Meyer.

Lörscher, W. 1983. Linguistische Beschreibung und Analyse von Fremdsprachenunterricht als Diskurs. Tübingen: Narr.

Mehan, H. 1979. Learning lessons: social organisation in the classroom. Cambridge (Mass.): Harvard University Press.

Meldau, R., and I. Pearcy. 1961. Deutsch-englisches Wörterbuch der Unterrichtssprache. München: Hueber.

Müller, M., and E. Zellmer. 1952. Englische Schulredensarten. Iserlohn: Silva.

Ostberg, H.K. 1982. Gesprächsverhalten in der Fremdsprache (Englisch) und fremdsprachlicher Unterricht. Augsburg

Priesemann, G. 1971. Zur Theorie der Unterrichtssprache. Düsseldorf: Schwann.

Sinclair, J.McH., and D. Brazil. 1982. Teacher talk. London: Oxford University Press.

Sinclair, J.McH., and R.M. Coulthard. 1975. Towards an analysis of discourse.London: Oxford University Press.

Weidner, L. 1968. Classroom phrases. München: Hueber.

Willis, J. 1981. Teaching English through English. London: Longman

Windmann, H. 1959. Englische Redewendungen. Münster: Aschendorff.